FLUID LIMITS FOR TIME-VARYING MANY-SERVER QUEUES WITH FINITE CAPACITY

MINGRUI WANG AND PRAKASH CHAKRABORTY

ABSTRACT. This paper develops fluid limits for nonstationary many-server loss systems with general service-time distributions. For the zero-buffer $M_t/G/n/n$ queuing model, we prove a functional strong law of large numbers for the fraction of busy servers and characterize the limit by a nonlinear Volterra integral equation with discontinuous coefficients induced by instantaneous blocking. Well-posedness is established through an appropriate solution concept, yielding the time-varying acceptance probability without heuristic approximations. We then treat the finite-buffer $M_t/G/n/(n+b_n)$ regime, proving a functional strong law of large numbers for the triplet of fractions of busy servers, occupied buffers, and cumulative departures, whose limit satisfies a coupled system of three discontinuous Volterra equations capturing the interaction of service completions, buffer occupancy, and admission control at the capacity boundary. We establish well-posedness and convergence of the time-varying acceptance probability. Our theoretical results are supported by numerical simulations for both zero and finite-buffer regimes, illustrating the convergence of transient acceptance probabilities guaranteed by our theory. Finally, we use the fluid limits to derive optimal staffing and buffer-capacity for both time-varying loss systems.

1. Introduction

Many modern service systems operate with limited capacity, meaning customers are turned away or lost when the system is full. Classic examples include telephone networks with a fixed number of trunk lines [11,18,21], hospital or emergency units with limited beds [1,2,9], wireless and optical networks with bandwidth and channel constraints [25,36,37,41], emergency services like ambulances and self-driving cars [17]. These loss models, sometimes called Erlang loss systems, have been studied extensively under steady-state conditions. In fact, the famous Erlang-B formula [12] developed over a century ago for telephone traffic gives the steady-state blocking probability for an M/M/n/n queue and remains a cornerstone result in stationary loss models. Yet real-world systems are rarely stationary: arrival rates and service demands fluctuate over time, and service durations are not necessarily memoryless. As a result, steady-state measures often fail to capture short-term dynamics, leading to inefficient or unstable operational decisions. Nonstationary, non-Markovian loss systems such as $M_t/G/n/n$ queues are significantly more challenging to analyze, and closed-form transient performance formulas are virtually impossible to obtain. This difficulty motivates the use of stochastic-process approximations for performance analysis, especially in many-server regimes where the number of servers n is large.

Fluid limits or functional strong laws of large numbers (FSLLN) provide deterministic approximations to many-server queuing systems by tracking the scaled system state as $n \to \infty$. These limits reveal the macroscopic law of motion of complex stochastic systems. Foundational work such as [16, 28] introduced asymptotic techniques for many-server systems and Markovian service networks. Subsequent research established fluid and diffusion limits under increasingly general conditions, including time-varying arrivals and non-exponential service times [20,26,27,33,42]. In contrast to these limit theorems, an extensive applied literature has developed practical approximations and

Date: November 12, 2025.

P. Chakraborty is partially supported by the National Science Foundation under grant DMS-2153915.

staffing heuristics for time-varying service systems. Related work, including [14,15,19,38,39], proposes pointwise-stationary (POS), modified-offered-load (MOL), and other transient approximations aimed at dynamic staffing, capacity planning, and transient performance evaluation. These studies underscore the need for rigorous transient characterizations that connect operational heuristics with asymptotic theory.

1.1. Overview of Approach and Key Insights. This paper develops a rigorous fluid-limit framework for analyzing time-varying many-server loss systems. Specifically, we study a sequence of systems with nonhomogeneous Poisson process (NHPP) arrivals and general service-time distributions, where both the number of servers and the arrival rate scale linearly with system size. The resulting limit is characterized by a nonlinear Volterra integral equation (VIE) that captures the transient evolution of the system's occupancy and, crucially, yields the time-dependent blocking and acceptance probabilities in the large-scale regime. Our work builds upon [8], which established a fluid limit for the nonstationary many-server $M_t/G/n/n$ loss system using a semimartingale representation of the instantaneous acceptance mechanism. We enhance that framework by introducing a refined convergence proof based on the discontinuous Volterra equation methodology of [22], ensuring well-posedness and uniqueness of the limit even under nonsmooth boundary dynamics.

A distinguishing feature of our analysis is the emergence of nonlinear Volterra equations with discontinuous coefficients, induced by the instantaneous blocking constraint at full capacity. This structure departs sharply from classical Markovian formulations and provides a new analytic mechanism to capture threshold-type, transient blocking phenomena in nonstationary systems. The discontinuity is not merely a technical complication. It serves as the deterministic counterpart of the system's stochastic acceptance barrier and encodes the operational behavior of loss systems under time-varying load.

From a methodological standpoint, our results bridge three traditions in the study of nonstationary queues: (i) steady-state or quasi-stationary approximations such as the Erlang-B, PSA, and MOL methods [14, 15, 29, 40]; (ii) computational and moment-based approximation methods, including cumulant and truncated-ODE approaches for time-varying loss and many-server systems [19, 30, 31, 38, 39]; and (iii) rigorous asymptotic limit theorems [16, 28, 33]. The fluid model derived here serves simultaneously as a limit theorem and a computational engine. It is a deterministic equation directly solvable by numerical methods, providing transient blocking probabilities without Monte Carlo simulations. This connection between rigorous scaling limits and practical performance computation strengthens the link between applied probability and operational analysis, particularly in time-dependent service environments such as healthcare scheduling, cloud service provisioning, and mobility-on-demand platforms. Accurate transient blocking or acceptance probabilities support dynamic staffing and admission-control decisions under fluctuating demand. Whereas traditional time-varying approximations assume local equilibrium, our limit provides a theoretically consistent foundation for approximating time-varying acceptance probabilities under nonstationary demand, which is central to time-dependent operations management.

Although our analysis focuses on NHPP arrivals, the fluid-limit structure depends only on the arrival-rate trajectory rather than Poisson-specific properties. The same analytical framework extends naturally to renewal or Cox processes with time-dependent intensities. This generality implies that the derived acceptance probabilities provide accurate first-order approximations for a broad class of time-varying queuing systems, highlighting the structural robustness of the fluid-limit formulation.

1.2. Contributions. We establish functional strong laws of large numbers for nonstationary manyserver loss systems under general service-time distributions. Both the zero-buffer $(M_t/G/n/n)$ and finite-buffer $(M_t/G/n/(n+b_n))$ systems are analyzed in a common framework that scales the number of servers and the arrival rate proportionally with system size ¹. The resulting limits are deterministic trajectories described by nonlinear Volterra integral equations (VIEs) that capture the transient evolution of the system occupancy and yield the associated time-dependent blocking and acceptance probabilities.

(i) Zero-buffer systems. For the $M_t/G/n/n$ model with nonhomogeneous Poisson arrivals of rate $\lambda(\cdot)$ and i.i.d. service times with distribution G, let N_t^n denote the number in system and $\bar{N}_t^n = N_t^n/n$ its scaled occupancy. We prove that when the system starts empty \bar{N}^n converges almost surely to a deterministic function ρ (see Theorem 3.3 for a more general and precise formulation) such that ρ solves the following discontinuous VIE:

$$\rho_t = \int_0^t \mathbb{1}_{\{\rho_{u-} < 1\}} \bar{G}(t - u) \lambda(u) du. \tag{1.1}$$

where \bar{G} is the service-time survival function. The integral equation above has discontinuous coefficients due to the indicator $\mathbb{1}_{\{\rho_u < 1\}}$, reflecting instantaneous blocking at capacity. We refine the convergence analysis of [8] by introducing the discontinuous Volterra solution concept of [22]. Specifically, ρ solves (1.1) if there exists an auxiliary acceptance function $w(\cdot)$ such that

$$\rho_t = \int_0^t w(u)\bar{G}(t-u)\lambda(u)du.$$

which ensures well-posedness even under nonsmooth boundary dynamics. As a corollary (see Corollary 3.1 for a precise formulation), we obtain that for λ -almost every t, the acceptance probability

$$P(\bar{N}_t^n < 1) \to w(t). \tag{1.2}$$

In addition, we identify $w(t) = \frac{d(t)}{\lambda(t)} \wedge 1$, where d(t) is the instantaneous departure rate, which agrees with heuristic expectations. Thus our analysis yields a rigorous FSLLN that provides a direct functional relationship between the time-varying acceptance (or blocking) probability and the system primitives through a deterministic limit equation (1.2).

(ii) Finite-buffer systems. We extend the analysis to the $M_t/\bar{G}/n/(n+b_n)$ model, where the buffer size b_n may scale with n so that $b_n/n \to \beta \in [0, \infty)$. Denote by \bar{S}_t^n , \bar{Q}_t^n , and \bar{D}_t^n the scaled numbers of busy servers, queued customers, and cumulative departures, respectively. We prove (see Theorem 4.3 for a more precise formulation) that the joint limit of these processes $(\bar{S}^n, \bar{Q}^n, \bar{D}^n)$ is given by the tuple (ρ, η, D) that satisfies a system of three coupled nonlinear VIEs:

$$\rho_{t} = \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}} \bar{G}(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}} \bar{G}(t-u)d(u)du,$$

$$\eta_{t} = \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}=1\}} \mathbb{1}_{\{\eta_{u-}<\beta\}} \lambda(u)du - \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}} d(u)du,$$

$$D_{t} = \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}} G(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}} G(t-u)d(u)du,$$

where $d(\cdot)$ denotes the fluid departure rate. These equations jointly describe the evolution of service completions, queue occupancy, and admission control at the boundary. As in the zero-buffer case, they are interpreted through auxiliary acceptance functions (w^1, w^2, w^3) ensuring existence and uniqueness of the limit. The resulting acceptance probability satisfies

$$P(\bar{Q}_t^n < \frac{b_n}{n}) \to w^3(t),$$

¹our analysis readily extends to time-varying piecewise constant service and buffer capacities. However, for simplicity, we consider the case where both are constant.

where $w^3(t) = \frac{d(t)}{\lambda(t)} \wedge 1$ as in the zero-buffer case. This extension introduces significant technical challenges beyond the zero-buffer case, requiring new arguments to handle the emerging coupled nonlinear Volterra systems.

(iii) Analytical and operational significance. The discontinuous Volterra framework developed here provides the first rigorous characterization of transient blocking and acceptance probabilities in large-scale, time-varying service systems with general service-time distributions. It yields a numerically tractable representation: the limit equations can be solved efficiently via numerical methods, enabling direct computation of transient performance measures without simulation. Beyond analytical clarity, the framework serves as a practical foundation for operational decision-making. We demonstrate its use for optimal staffing and buffer capacity design, showing how the deterministic fluid model can approximate system-level performance with high accuracy. These formulations extend naturally to dynamic versions, where time-dependent staffing or capacity policies adapt to fluctuating demand. Overall, these results unify the transient analysis of Erlang loss and delay systems and offer a theoretically grounded computational tool for performance evaluation and dynamic control in applications such as call centers, hospitals, and cloud-service platforms.

Together, the zero and finite-buffer results form an integrated theory of time-varying many-server systems. The discontinuous Volterra formulation opens the door to higher-order diffusion refinements and control-theoretic extensions.

1.3. Paper Organization. The remainder of the paper is organized as follows. Section 2 presents the preliminaries, including notation, key probability results, weak convergence tools, and the analytical framework for discontinuous Volterra integral equations (VIEs). Section 3 focuses on the zero-buffer $M_t/G/n/n$ system. We derive the fluid limit, prove the functional strong law of large numbers, and establish convergence of the time-varying acceptance and blocking probabilities. Section 4 extends the analysis to the finite-buffer $M_t/G/n/(n+b_n)$ model. Here, we characterize the joint fluid limit of the fractions of busy servers, occupied buffers, and departures as the solution to a system of coupled Volterra integral equations, and we prove convergence of the corresponding acceptance and blocking probabilities. Section 5 provides numerical experiments that illustrate the accuracy and interpretability of the fluid-limit approximation across both zero- and finite-buffer regimes, in addition to optimal server and capacity applications. A brief concluding Section 6 summarizes the findings and outlines potential extensions, including diffusion refinements and control applications.

2. Preliminaries and Notations

In this section we present some preliminary results that will be useful later on.

2.1. Convergence in Skorokhod Space. Let $\mathbb{D} = \mathbb{D}[0,T]$ denote the space of càdlàg (right-continuous with left limits) functions on [0,T]. For a function $f \in \mathbb{D}$ and a set $T_0 \subseteq [0,T]$, we denote its modulus of continuity on T_0 as

$$w_f(T_0) = \sup_{s,t \in T_0} |f(t) - f(s)|.$$

For any $\delta \in (0,T)$, let

$$w_f'(\delta) = \inf_{\mathcal{P}: \|\mathcal{P}\| \le \delta} \max_{0 < i \le |\mathcal{P}|} w_f([t_{i-1}, t_i)),$$

where \mathcal{P} runs over the set of all partitions of [0,T], in the sense that a generic \mathcal{P} looks like

$$\mathcal{P} = \left\{ 0 = t_0, \dots, t_{|\mathcal{P}|} = T \right\},\,$$

and $\|\mathcal{P}\|$ denotes the mesh or norm of the partition \mathcal{P} :

$$\|\mathcal{P}\| = \max_{1 \le i < |\mathcal{P}|} |t_i - t_{i-1}|.$$

A function f belongs to the space \mathbb{D} if and only if

$$\lim_{\delta \downarrow 0} w_f'(\delta) = 0.$$

For a proof and related discussion, see [3, Chapter 13]. The Skorokhod distance between two functions $f, g \in \mathbb{D}$ is defined as

$$d_S(f,g) = \inf_{\lambda \in \Lambda} \max \left\{ \sup_{t \in [0,T]} |\lambda(t) - t|, \sup_{t \in [0,T]} |f(\lambda(t)) - g(t)| \right\},$$

where Λ is the class of strictly increasing, continuous mappings of [0,T] to itself. The topology on $\mathbb D$ induced by this metric is known as the Skorokhod topology. It can be shown that $\mathbb D$ is not a complete space with respect to the Skorokhod distance d_S but there exists a topologically equivalent metric d_0 with respect to which $\mathbb D$ is complete. For $0 \le t_1 < \cdots < t_k \le T$, define the natural projection $\pi_{t_1 \cdots t_k}$ from $\mathbb D$ to $\mathbb R^k$ as:

$$\pi_{t_1\cdots t_k}(x) = (x(t_1), \dots x(t_k)),$$

and the Borel σ -field of \mathbb{D} as \mathcal{D} . For probability measures \mathbb{P} on $(\mathbb{D}, \mathcal{D})$, denote by $T_{\mathbb{P}}$ the set of t in [0, T] for which the projection π_t is continuous except at points forming a set of \mathbb{P} -measure 0. We include some useful results from [3]:

Theorem 2.1. A sequence of probability measures $\{\mathbb{P}_n\}$ on $(\mathbb{D}, \mathcal{D})$ is tight if and only if:

$$\lim_{a \to \infty} \limsup_{n} \mathbb{P}_n \left[x : \sup_{t \in [0,T]} |x(t)| \ge a \right] = 0,$$

and for each $\varepsilon > 0$,

$$\lim_{\delta} \lim_{n} \sup_{n} \mathbb{P}_{n} \left[x : w'_{x}(\delta) \geq \varepsilon \right] = 0.$$

Theorem 2.2. If $\{\mathbb{P}_n\}$ is tight, and if $\mathbb{P}_n\pi_{t_1\cdots t_k}^{-1} \Rightarrow \mathbb{P}\pi_{t_1\cdots t_k}^{-1}$ holds whenever $t_1, \ldots t_k$ all lie in $T_{\mathbb{P}}$, then $\mathbb{P}_n \Rightarrow \mathbb{P}$.

2.2. Counting Measure. Let $(\Omega, \mathcal{F}, \mathcal{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space. Let $(N_t)_{t\geq 0}$ be a point process given by a sequence $(T_n)_{n\geq 1}$ of jump times, that is

$$N_t := N((0, t]) = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}},$$

where $N(\cdot) = \sum_{n\geq 1} \delta_{T_n}$ is the corresponding counting measure and δ_y stands for the Dirac measure at y. Suppose in addition the n^{th} jump time or arrival T_n has a corresponding mark or random variable Z_n taking values in some measurable space (E, \mathcal{E}) . Then $(T_n, Z_n)_{n\geq 1}$ is called an E-marked point process. Let $\mathcal{M}^N(\cdot \times \cdot)$ be the counting measure of the marked point process, that is, for each $C \in \mathbb{R}, L \in \mathcal{E}$

$$\mathcal{M}^{N}(C \times L) = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_{i} \in C\}} \mathbb{1}_{\{Z_{i} \in L\}}.$$

This implies for measurable functions $\varphi:(\mathbb{R},\mathcal{B}(\mathbb{R}))\times(E,\mathcal{E})\to(\overline{\mathbb{R}},\mathcal{B}(\overline{\mathbb{R}}))$

$$\int_0^t \int_E \varphi(u, z) \mathcal{M}^N(du \times dz) = \sum_{i=1}^\infty \varphi(T_i, Z_i) \mathbb{1}_{\{T_i \le t\}}.$$

We recall the notions of intensity measure and intensity function following [5].

Definition 2.1. [5, Def 10.2.13] The intensity measure ν of a locally finite point process N on \mathbb{R}^m is defined by

$$C \mapsto \nu(C) := \mathbb{E}[N(C)] \quad (C \in \mathcal{B}(\mathbb{R}^m)).$$

In addition, if ν is of the form $\nu(C) = \int_C \zeta(x) dx$ for some non-negative measurable function ζ : $\mathbb{R}^m \to \mathbb{R}$, the point process N is said to admit the intensity function $\zeta(x)$.

For a point process N with intensity measure ν and intensity function ζ , we introduce the Campbell's formula from [5, Thm 10.2.15]:

Theorem 2.3. For all measurable functions $\varphi : \mathbb{R}^m \to \mathbb{R}$ which are non-negative or ν -integrable, the integral $\int_{\mathbb{R}^m} \varphi(x) N(dx)$ is well defined and

$$\mathbb{E}\left[\int_{\mathbb{R}^m}\varphi(x)N(dx)\right] = \int_{\mathbb{R}^m}\varphi(x)\nu(dx) = \int_{\mathbb{R}^m}\varphi(x)\zeta(x)dx.$$

In particular, $\int_{\mathbb{R}^m} \varphi(x) N(dx)$ is a.s. finite if φ is ν -integrable.

2.3. **Discontinuous Volterra Integral Equation.** We recall the notion of solution for discontinuous Volterra integral equations, as presented in [22]. First, we introduce some related notations. For any $p \in L^{\infty}_{loc}(-\infty, \infty)$ and any $\epsilon > 0$, define:

$$\underline{p}_{\epsilon}(t) = \underset{|t-s| < \epsilon}{\operatorname{ess inf}} p(s), \quad \bar{p}_{\epsilon}(t) = \underset{|t-s| < \epsilon}{\operatorname{ess sup}} p(s).$$

In addition, for $t \in [0, T]$ define:

$$\underline{p}(t) = \lim_{\epsilon \to 0} \underline{p}_{\epsilon}(t), \quad \bar{p}(t) = \lim_{\epsilon \to 0} \bar{p}_{\epsilon}(t). \tag{2.1}$$

Definition 2.2. Let $p:[0,\infty)\to\mathbb{R}$ and $q:[0,T]\to\mathbb{R}$ be bounded functions. Furthermore, let $a\in L^1[0,T]$. A pair of functions $x:[0,T]\to\mathbb{R}$ and $z:[0,T]\to\mathbb{R}$ is said to be a solution of the Volterra integral equation

$$x(t) + \int_0^t a(t-s)p(x(s))ds = q(t), \quad 0 \le t \le T$$

if x and z are bounded and

$$p(x(t)) \le z(t) \le \bar{p}(x(t))$$
 a.e., $0 \le t \le T$,

such that

$$x(t) + \int_0^t a(t-s)z(s)ds = q(t), \quad 0 \le t \le T.$$

Remark 2.1. We point out to the reader that the assumption on p,q and a can be relaxed or modified as done in [23,24]. Our exposition here is chosen for simplicity and the specific processes we encounter later.

- 2.4. **Notations.** We employ the following notations for different modes of convergence:
 - $\stackrel{p}{\rightarrow}$: Convergence in probability of random variables or stochastic processes,
 - ⇒: Weak convergence for probability measures or random variables,
 - *: Weak-star convergence in general function spaces,
 - $\stackrel{\mathbb{D}}{\rightarrow}$: Convergence in the Skorokhod topology.

3. Fluid limit for zero-buffer loss system

3.1. **Setup.** In this section, we introduce the zero-buffer loss queuing model. We consider a sequence of queuing systems indexed by n, subject to the following assumptions.

Assumption 3.1. Consider a $M_t/G/n/n$ loss queuing system; namely, a queuing system with

- i. a nonhomogeneous Poisson arrival process A^n with rate or intensity function $n\lambda(\cdot)$, where λ is locally integrable;
- ii. general customer service times sampled independently from a distribution G with density g;
- iii. the system has n servers and zero buffer or waiting space. That is, when all n servers are busy, incoming customer arrivals are lost. Equivalently, the customers can be thought to have 0 patience.

Remark 3.1. Note that the intensity function corresponding to the arrival process A^n could be extended to a more general λ_n for all n, such that $\lambda_n/n \to \lambda$ under some topology. This generalization should be an easy extension and not considered in this article to keep considerations simpler.

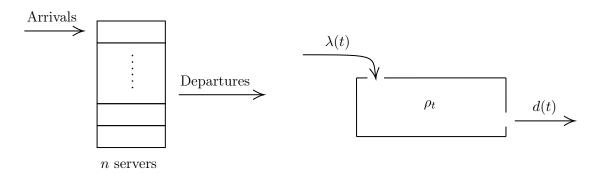


Figure 1. Zero-buffer loss system and its fluid model

3.2. Fraction of Occupied Servers or Scaled Number in System. Consider the $M_t/G/n/n$ loss system as in Assumption 3.1. Let T_i and V_i represent the arrival and service times, respectively, of the i-th customer. Let N_t^n denote the number of occupied servers, or equivalently the number of customers in the system at time t. Denote $\bar{N}_t^n := \frac{N_t^n}{n}$ to be the fraction of occupied servers or the n-scaled number of customers in the system at time t. Also, let \mathcal{F}_t^n be the filtration generated by $\{\bar{N}_s^n : s \in [0,t]\}$.

For the sake of simplicity, we first assume that the system starts empty, that is, the number of customers in the system at time 0 is zero. In the sequel, we will relax this assumption.

Observe that the number of busy servers at time t consists of all arrivals to the system such that all of the following conditions are met:

- (i) the customer arrival occurs at or prior to time t,
- (ii) the number of occupied servers upon the customer's arrival is less than n, and
- (iii) the remaining service time of this customer at time t is positive, that is, the customer is yet to depart the system.

For the *i*-th customer arriving to the system, these conditions correspond to $\{T_i \leq t\}$, $\{N_{T_i}^n < n\}$ or $\{\bar{N}_{T_i}^n < 1\}$, and $\{V_i > t - T_i\}$ respectively. Consequently, the number of customers at time t satisfies

$$N_t^n = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{N_{T_i}^n < n\}} \mathbb{1}_{\{V_i > t - T_i\}}.$$
(3.1)

On scaling (3.1) by n, we obtain that the fraction of occupied servers satisfies

$$\bar{N}_t^n = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{\bar{N}_{T_i}^n < 1\}} \mathbb{1}_{\{V_i > t - T_i\}}.$$
(3.2)

Crucially, observe that N^n or \bar{N}^n given by (3.1)-(3.2) are given by integral equations whose evolution depends, in general, on its history. As such, these processes are non-Markovian and in this work we provide a way of obtaining scaling limits of such processes arising out of loss queuing systems. To that effect, we work with the scaled process \bar{N}^n and obtain a representation using random measures. Denote

$$W_n(t, u, x) = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{\bar{N}_{u-}^n < 1\}} \mathbb{1}_{\{x > t - u\}} \mathbb{1}_{\{u \le t\}}.$$
 (3.3)

Then relation (3.2) can be represented as

$$\bar{N}_t^n = \int_0^t \int_{\mathbb{R}} W_n(t, u, x) \mathcal{M}^n(du, dx), \tag{3.4}$$

where \mathcal{M}^n is the counting measure associated with the marked point process of the arrival and service time pair (T_i, V_i) . Taking expectation, we have by Theorem 2.3 that

$$\mathbb{E}\left[\int_0^t \int_{\mathbb{R}} W_n(t,u,x) \mathcal{M}^n(du,dx)\right] = \int_0^t \int_{\mathbb{R}} W_n(t,u,x) n\lambda(u) g(x) du dx.$$

Denote \mathcal{M}_*^n to be the compensated random measure:

$$\mathcal{M}_*^n = \mathcal{M}^n - \mathcal{M}_c^n, \tag{3.5}$$

where $\mathcal{M}_{c}^{n}(du, dx) := \mathbb{E}\left[\mathcal{M}^{n}(du, dx)\right] = n\lambda(u)g(x)dudx$.

Having obtained an integral representation for the fraction of occupied servers in (3.4), our goal is to exploit this relation to obtain the limit of the stochastic process $\{\bar{N}_t^n, t \geq 0\}$ as n goes to infinity. We begin with a result proving convergence along a subsequence.

Proposition 3.1. Let Assumption 3.1 hold. Assume that the system starts empty, that is $\rho_0^n = 0$ for all n. Then

(i) For any T > 0 and any subsequence, there exists a further subsequence (r_k) and a continuous, possibly stochastic process ρ such that almost surely,

$$\bar{N}^{r_k} \to \rho,$$
 (3.6)

in the uniform topology.

(ii) Moreover, given (r_k) , almost surely there exists a bounded, possibly stochastic process w such that

$$\mathbb{1}_{\{\bar{N}_{t-}^{r_k} < 1\}} \stackrel{*}{\rightharpoonup} w(t) \quad in \ L^{\infty}[0, T]. \tag{3.7}$$

(iii) Furthermore, almost surely, ρ and w defined in (3.6)-(3.7) satisfy

$$\rho_t = \int_0^t w(u)\bar{G}(t-u)\lambda(u)du, \quad t \in [0,T], \quad and$$

$$\mathbb{1}_{\{\rho_{u-}<1\}} \le w(u) \le 1, \ a.e. \ in \ [0,T].$$
(3.8)

That is, for almost all $\omega \in \Omega$ $(\rho(\omega), w(\omega))$ as in (3.8) is a solution, interpreted according to Definition 2.2, to the following non-linear discontinuous Volterra integral equation

$$\rho_t = \int_0^t \mathbb{1}_{\{\rho_{u-}<1\}} \bar{G}(t-u)\lambda(u)du.$$
 (3.9)

Proof. For simplicity we will consider the initial subsequence to be (n), but the arguments below go through for any initial subsequence.

Part (i). Applying the decomposition (3.5) in (3.4) we have

$$\bar{N}_t^n = X_t^n + Y_t^n, \tag{3.10}$$

where

$$X_t^n := \int_0^t \int_{\mathbb{R}} W_n(t, u, x) \mathcal{M}_*^n(du, dx), \text{ and } Y_t^n := \int_0^t \mathbb{1}_{\{\bar{N}_{u-}^n < 1\}} \bar{G}(t - u) \lambda(u) du.$$
 (3.11)

We will analyze X^n and Y^n separately, starting with the term Y^n .

By the local integrability of λ from Assumption 3.1, we have from (3.11) that almost surely

$$\sup_{n} \sup_{t \in [0,T]} Y_t^n \le \int_0^T \lambda(u) \, du < \infty. \tag{3.12}$$

Meanwhile Y^n satisfies

$$Y_{t}^{n} - Y_{s}^{n} = \int_{0}^{t} \mathbb{1}_{\{\bar{N}_{u-}^{n} < 1\}} \bar{G}(t-u)\lambda(u)du - \int_{0}^{s} \mathbb{1}_{\{\bar{N}_{u-}^{n} < 1\}} \bar{G}(s-u)\lambda(u)du$$

$$= \int_{s}^{t} \mathbb{1}_{\{\bar{N}_{u-}^{n} < 1\}} \bar{G}(t-u)\lambda(u)du + \int_{0}^{s} \mathbb{1}_{\{\bar{N}_{u-}^{n} < 1\}} \left(\bar{G}(t-u) - \bar{G}(s-u)\right)\lambda(u)du. \tag{3.13}$$

Given that \bar{G} is non-increasing and bounded above by 1, we can derive from (3.13) that

$$\sup_{n} |Y_t^n - Y_s^n| \le \int_s^t \lambda(u) du.$$

Since the function $\Lambda(t) = \int_0^t \lambda(u) du$ is uniformly continuous on [0, T], it follows that Y^n are equicontinuous. Therefore we have

$$\lim_{\delta \downarrow 0} \sup_{n} w_{Y^n}'(\delta) = 0. \tag{3.14}$$

By (3.12), (3.14), Theorem 2.1 and Prokhorov's theorem we can conclude that there exists $\rho \in \mathbb{D}$ and a subsequence (n_k) such that

$$Y^{n_k} \stackrel{\mathbb{D}}{\to} \rho$$
, almost surely. (3.15)

Moreover, $L^1[0,T]$ is a separable Banach space with dual $L^{\infty}[0,T]$ and $\mathbb{1}_{\{\bar{N}_{u-}^n < 1\}} \in L^{\infty}[0,T]$. Therefore by [6, Thm 2.34], almost surely there is a subsubsequence $(l_k) \subset (n_k)$ and $w \in L^{\infty}[0,T]$, possibly depending on (l_k) , such that for any $\phi \in L^1[0,T]$

$$\lim_{k \to \infty} \int_0^t \phi(u) \mathbb{1}_{\{\bar{N}_{u^-}^{l_k} < 1\}} du = \int_0^t \phi(u) w(u) du, \quad \text{for all } t \in [0, T].$$
 (3.16)

Note that w could still be random at this stage. In particular, choosing $\phi(\cdot) = \bar{G}(t - \cdot)\lambda(\cdot)$ we have for all $t \in [0, T]$, almost surely

$$\lim_{k \to \infty} \int_0^t \mathbb{1}_{\{\bar{N}_{u-}^{l_k} < 1\}} \bar{G}(t-u)\lambda(u)du = \int_0^t w(u)\bar{G}(t-u)\lambda(u)du. \tag{3.17}$$

From (3.17) we identify ρ in (3.15), that is:

$$\rho_t = \int_0^t w(u)\bar{G}(t-u)\lambda(u)du. \tag{3.18}$$

This limiting function ρ is continuous because \bar{G} and w are bounded, and λ is integrable. It follows that the convergence in (3.15) is also under the uniform topology:

$$\lim_{k \to \infty} \sup_{t \in [0,T]} \left| Y_t^{l_k} - \rho_t \right| = 0, \quad \text{almost surely.}$$
 (3.19)

Let us now analyze the term X^n . By (3.10)-(3.12) we have that almost surely

$$\sup_{n} \sup_{t \in [0,T]} |X_{t}^{n}| \le \sup_{n} \sup_{t \in [0,T]} \bar{N}_{t}^{n} + \sup_{n} \sup_{t \in [0,T]} Y_{t}^{n} \le 1 + \int_{0}^{T} \lambda(u) \, du < \infty. \tag{3.20}$$

Furthermore, the number of jumps of \bar{N}^n is bounded by twice that of the arrivals. Consequently \bar{N}^n is piecewise constant with almost surely finitely many jumps in [0,T]. Thus we have for all n,

$$w'_{\bar{N}^n}(\delta) = 0, (3.21)$$

almost surely. Using (3.21) and (3.14) in relation (3.10) we get

$$\lim_{\delta \downarrow 0} \sup_{n} w'_{X^n}(\delta) = 0. \tag{3.22}$$

By (3.20), (3.22) and Theorem 2.1 we obtain the tightness of $(X^n)_{n\geq 1}$. Now recalling W_n in (3.3) and X^n in (3.11), we have for any fixed $t\in [0,T]$

$$\mathbb{E}(X_t^n)^2 = \mathbb{E}\left[\int_0^t \int_{\mathbb{R}} W_n(t, u, x) \mathcal{M}_*^n(du, dx)\right]^2$$

$$\leq \frac{1}{n^2} \mathbb{E}\left[\int_0^t \int_{\mathbb{R}} \mathcal{M}_*^n(du, dx)\right]^2 = \frac{1}{n^2} \operatorname{Var}(A_t^n) = \frac{1}{n} \int_0^t \lambda(u) du \to 0,$$

as $n \to \infty$. Consequently for each $t \in [0,T], X_t^n \xrightarrow{p} 0$. Thus for any $(t_1,t_2,\ldots,t_d) \in [0,T]^d$, the finite dimensional vectors $(X_{t_1}^n,\ldots,X_{t_d}^n) \xrightarrow{p} (0,\ldots,0)$ as a consequence of the Cramer-Wold device [4, Thm 29.4]. By Theorem 2.2 we have that X^n converges in distribution to the constant zero function. Since the limiting function is non-random, the convergence becomes:

$$X^n \xrightarrow{p} 0$$
, in the uniform topology. (3.23)

From (3.23) we know that there exists a subsequence $(r_k) \subset (l_k)$, such that

$$\sup_{t \in [0,T]} |X_t^{r_k}| \to 0, \quad \text{almost surely.}$$
 (3.24)

Combining the above arguments together, for this sequence (r_k) , we thus obtain from (3.10), (3.19) and (3.24) that almost surely

$$\bar{N}^{r_k} = X^{r_k} + Y^{r_k} \to \rho, \tag{3.25}$$

in the uniform topology, where ρ is identified by (3.18). This completes the proof of Part (i).

Part (ii). This has already been shown above in (3.16).

Part (iii). The function ρ has been identified by (3.18). It remains to show constraint for the function w. We now observe that for sequence (\bar{N}^n) the set $\{\bar{N}^n_{u-} < 1\}$ is identical to the set $\{\bar{N}^n_{u-} \le 1 - \frac{1}{n}\}$. This is because \bar{N}^n only takes values in $\{\frac{i}{n}: i = 1, \ldots, n\}$. Therefore, we can rewrite (3.10) as

$$\bar{N}_t^n = X_t^n + \int_0^t \mathbb{1}_{\{\bar{N}_{u-}^n \le 1 - 1/n\}} \bar{G}(t-u) \lambda(u) du.$$

Notice that by Proposition 3.1, $\rho \leq 1$. Our next objective is to discover the function w in (3.8). Since ρ is continuous, fix $\varepsilon > 0$ and choose N large enough such that for all k > N we have $r_k > \frac{3}{\varepsilon}$, and $\|\bar{N}^{r_k} - \rho\|_T < \frac{\varepsilon}{3}$ almost surely. Then it is readily checked that

$$\mathbb{1}_{\{\rho_{u-} \le 1 - \varepsilon\}} \le \mathbb{1}_{\{\bar{N}_{u-}^{r_k} \le 1 - 1/r_k\}} \le \mathbb{1}_{\{\rho_{u-} < 1 + \varepsilon\}}.$$

Therefore for any $\phi \geq 0$ such that $\phi \in L^1[0,T]$ we have almost surely

$$\int_0^t \phi(u) \mathbb{1}_{\{\rho_u - \le 1 - \varepsilon\}} du \le \int_0^t \phi(u) \mathbb{1}_{\{\bar{N}_{u-}^{r_k} \le 1 - 1/r_k\}} du \le \int_0^t \phi(u) \mathbb{1}_{\{\rho_u - < 1 + \varepsilon\}} du.$$

Note that $\lim_{\varepsilon \downarrow 0} \mathbb{1}_{\{\rho_{u-} < 1 - \varepsilon\}} = \mathbb{1}_{\{\rho_{u-} < 1\}}$ and $\lim_{\varepsilon \downarrow 0} \mathbb{1}_{\{\rho_{u-} < 1 + \varepsilon\}} = \mathbb{1}_{\{\rho_{u-} \leq 1\}} = 1$. Consequently taking $k \to \infty$ and then $\varepsilon \downarrow 0$ we have by the dominated convergence theorem and (3.7) that almost surely:

$$\int_{0}^{t} \phi(u) \mathbb{1}_{\{\rho_{u-} < 1\}} du \le \int_{0}^{t} w(u) \phi(u) du \le \int_{0}^{t} \phi(u) du.$$

Since ϕ is arbitrary in $L^1[0,T]$, we have almost surely

$$\mathbb{1}_{\{\rho_{u-}<1\}} \le w(u) \le 1$$
, a.e. in $[0,T]$.

Recall the notations defined in (2.1). It is easily checked that $\underline{\mathbb{1}}_{\{\rho_u < 1\}} = \mathbb{1}_{\{\rho_u < 1\}}$ and $\overline{\mathbb{1}}_{\{\rho_u < 1\}} = \mathbb{1}_{\{\rho_u <$

We have established a fluid limit for \bar{N}_t^n along a subsequence when the system starts empty. Now, we extend our considerations to a more general case.

Assumption 3.2. Let the conditions under Assumption 3.1 hold. In addition let the number of customers in the system at time 0: N_0^n , satisfy

$$\lim_{n\to\infty}\frac{N_0^n}{n}=\rho_0, \quad almost \ surely,$$

where $\rho_0 \in [0,1]$. Moreover, assume that the remaining service times of each of the initially occupied servers follow the distribution F^n satisfying

$$\lim_{n \to \infty} \sup_{t} |F^{n}(t) - F(t)| = 0,$$

for some limiting distribution F.

Proposition 3.2. Let Assumption 3.2 hold. Then

(i) For any T > 0 and any subsequence of \bar{N}^n , there exists a further subsequence \bar{N}^{r_k} and a real-valued continuous, possibly stochastic process ρ such that almost surely,

$$\bar{N}^{r_k} \to \rho,$$
 (3.26)

in the uniform topology.

(ii) Moreover, given (r_k) , almost surely there exists a bounded, possibly stochastic process w such that

$$\mathbb{1}_{\{\bar{N}_{t-}^{r_{k}}<1\}} \stackrel{*}{\rightharpoonup} w(t) \quad in \ L^{\infty}[0,T]. \tag{3.27}$$

(iii) Furthermore, almost surely, ρ and w defined in (3.26)-(3.27) satisfy

$$\rho_t = \rho_0 \bar{F}(t) + \int_0^t w(u) \bar{G}(t - u) \lambda(u) du, \quad t \in [0, T], \quad and$$

$$\mathbb{1}_{\{\rho_u = <1\}} \le w(u) \le 1, \quad a.e. \text{ in } [0, T].$$
(3.28)

That is, for almost all $\omega \in \Omega$ $(\rho(\omega), w(\omega))$ as in (3.28) is a solution, interpreted according to Definition 2.2, to the following non-linear discontinuous Volterra integral equation

$$\rho_t = \rho_0 \bar{F}(t) + \int_0^t \mathbb{1}_{\{\rho_{u-} < 1\}} \bar{G}(t-u) \lambda(u) du.$$
 (3.29)

Proof. At time 0, the number of customers in service is N_0^n . Let the remaining service times for the customers in service be $(V_i^0)_{1 \le i \le N_0^n}$. Then, similar to (3.4) we have:

$$\bar{N}_t^n = \frac{1}{n} \sum_{i=1}^{N_0^n} \mathbb{1}_{\{V_i^0 > t\}} + \int_0^t \int_{\mathbb{R}} W_n(t, u, x) \mathcal{M}^n(du, dx).$$
 (3.30)

Observe that

$$\frac{1}{n} \sum_{i=1}^{N_0^n} \mathbb{1}_{\{V_i^0 > t\}} = \frac{N_0^n}{n} \frac{1}{N_0^n} \sum_{i=1}^{N_0^n} \mathbb{1}_{\{V_i^0 > t\}}.$$
(3.31)

By Assumption 3.2, thanks to Glivenko-Cantelli theorem

$$\lim_{n\to\infty}\sup_t\left|\frac{1}{N_0^n}\sum_{i=1}^{N_0^n}\mathbb{1}_{\{V_i^0>t\}}-\bar{F}(t)\right|=0,\quad\text{almost surely}.$$

Therefore from the decomposition (3.31) we have

$$\lim_{n \to \infty} \sup_{t} \left| \frac{1}{n} \sum_{i=1}^{N_0^n} \mathbb{1}_{\{V_i^0 > t\}} - \rho_0 \bar{F}(t) \right| = 0, \quad \text{almost surely.}$$
 (3.32)

Since we already analyzed the second term in (3.30) involving integration with respect to \mathcal{M}^n in Proposition 3.1, we obtain our desired result from (3.32).

Now, we establish the existence of a unique ρ that satisfies (3.29) in the sense of Definition 2.2. Consequently, we obtain a unique fluid limit of the fraction of occupied servers \bar{N}_t^n .

Theorem 3.1. Let Assumption 3.2 hold. Then there exists a unique solution ρ to the discontinuous Volterra integral equation (3.29). That is, there exists a unique solution ρ such that for all $t \in [0, T]$

$$\rho_t = \rho_0 \bar{F}(t) + \int_0^t z(u)\bar{G}(t-u)\lambda(u)du, \quad such \ that \ 0 \le \rho_t \le 1, \tag{3.33}$$

for some z(t) that satisfies

$$\mathbb{1}_{\{\rho_t < 1\}} \le z(t) \le 1 \quad a.e. \ in [0, T]. \tag{3.34}$$

Proof. The existence of the solution directly follows from Proposition 3.2. In order to prove uniqueness, let us define

$$\sigma_0 = 0, \ \tau_i = \inf_{t > \sigma_{i-1}} \{ t : \rho_t = 1 \} \text{ and } \sigma_i = \inf_{t > \tau_i} \{ t : \rho_t < 1 \}.$$
 (3.35)

We first show that there are at most countably many τ_i, σ_i . Denote \mathcal{I} the index set of τ_i, σ_i . Since ρ_t is continuous, by definition we know that $\tau_{i+1} > \sigma_i$. That is $\rho_t < 1$ for $t \in (\sigma_i, \tau_{i+1})$. Additionally, $\{(\sigma_i, \tau_{i+1})\}_{i \in \mathcal{I}}$ are pairwise disjoint open intervals on \mathbb{R} . Since each nonempty open interval in \mathbb{R} contains a rational, we can construct an injection $\mathcal{I} \to \mathbb{Q}$ to conclude \mathcal{I} is a countable set.

We will prove uniqueness by contradiction. Suppose there exist two solutions $(\rho_t^1, z_1(t))$ and $(\rho_t^2, z_2(t))$ satisfying (3.33) such that $\rho_t^1 \neq \rho_t^2$ for some $t \in [0, T]$. Denote

$$\sigma_0^1 = 0, \ \tau_i^1 = \inf_{t > \sigma_i^1} \{t : \rho_t^1 = 1\} \text{ and } \sigma_i^1 = \inf_{t > \tau_i^1} \{t : \rho_t^1 < 1\},$$

and similarly τ_i^2, σ_i^2 for ρ^2 , respectively. Since by (3.34) we have for $t \in \{s : \rho_s < 1\}$, z(t) = 1 is the only choice, we can conclude that the first time ρ_t^1 differs from ρ_t^2 can only be one of those σ_i^1, σ_i^2 . Define

$$i_0 = \min\{i \in \mathcal{I} \mid \sigma_i^1 \neq \sigma_i^2\}.$$

Since the index set \mathcal{I} is countable, and \mathbb{N} is well-ordered, the above term is well defined. Without loss of generality we can assume $\sigma_{i_0}^1 < \sigma_{i_0}^2$. Then, for $t \in [0, \sigma_{i_0}^1]$, we have

$$\rho_t^1 = \rho_0 \bar{F}(t) + \int_0^t z_1(u) \bar{G}(t-u) \lambda(u) du$$

= $\rho_0 \bar{F}(t) + \int_0^t z_2(u) \bar{G}(t-u) \lambda(u) du = \rho_t^2.$

Consequently for $t \in [0, \sigma_{i_0}^1]$

$$\int_0^t (z_1(u) - z_2(u)) \,\bar{G}(t - u)\lambda(u) du = 0.$$
(3.36)

Notice that

$$\frac{\partial}{\partial t} \left(z_1(u) - z_2(u) \right) \lambda(u) \bar{G}(t-u) = -\left(z_1(u) - z_2(u) \right) \lambda(u) g(t-u).$$

Since z_1, z_2 are bounded and $\lambda, g \in L^1[0, T]$, by Young's convolution inequality the function $(u, t) \mapsto (z_1(u) - z_2(u))\lambda(u)g(t-u) \in L^1([0, T] \times [0, T])$. Therefore, we can apply [35, Thm 2.7] to take derivatives of both side of (3.36) to obtain

$$(z_1(t) - z_2(t)) \lambda(t) - \int_0^t (z_1(u) - z_2(u)) \lambda(u) g(t - u) du = 0, \text{ a.e. in } [0, T].$$
 (3.37)

Now observe that the only solution in $L^1[0,T]$ to the Volterra integral equation

$$x(t) = \int_0^t x(u)g(t-u)du$$

is $x(t) \equiv 0$ (see for example [7, Thm 1.2.8]). Therefore, from (3.37) we have for $t \in [0, \sigma_{i_0}^1]$

$$(z_1(t) - z_2(t)) \lambda(t) = 0$$
, a.e. in $[0, T]$. (3.38)

Recall that $\sigma_{i_0}^1 < \sigma_{i_0}^2$. This implies by continuity of ρ^1 that there exists δ with $0 < \delta < \sigma_{i_0}^2 - \sigma_{i_0}^1$ such that

$$\rho_t^1 < \rho_t^2 = 1, \quad \text{for } t \in (\sigma_{i_0}^1, \sigma_{i_0}^1 + \delta).$$
 (3.39)

Therefore, from (3.33) we have

$$\rho_0 \bar{F}(t) + \int_0^t z_1(u) \lambda(u) \bar{G}(t-u) du < \rho_0 \bar{F}(t) + \int_0^t z_2(u) \lambda(u) \bar{G}(t-u) du, \quad \text{for } t \in (\sigma_{i_0}^1, \sigma_{i_0}^1 + \delta).$$

Plugging (3.38) into the above inequality we obtain

$$\int_{\sigma_{i_0}^1}^t (z_1(u) - z_2(u)) \,\lambda(u) \bar{G}(t-u) du < 0, \quad \text{for } t \in (\sigma_{i_0}^1, \sigma_{i_0}^1 + \delta).$$

Since $\lambda(u)\bar{G}(t-u) \geq 0$, there exists a positive measure set $\mathcal{A} \subset (\sigma_{i_0}^1, \sigma_{i_0}^1 + \delta)$ such that $u \in \mathcal{A}$ implies $z_1(u) - z_2(u) < 0$. However, by (3.34) and (3.39) we have for almost every $t \in (\sigma_{i_0}^1, \sigma_{i_0}^2)$, $z_1(u) = 1 \geq z_2(u)$. This is a contradiction. Therefore, ρ_t is unique.

In Theorem 3.1 we established the unique solvability of (3.33)-(3.34) in the sense that ρ_t is unique. Therefore, by Proposition 3.2 the fraction of occupied servers \bar{N}_t^n converge to this unique ρ_t . By (3.34) z(t)=1 when $\rho_t<1$. However, z(t) remains unspecified when $\rho_t=1$. It would be beneficial to specify a possible value of z(t) in this regime, specifically for the purpose of numerical experimentations. The following theorem provides a solution of z.

Theorem 3.2. Under the setting of Theorem 3.1, the pair (ρ, z) satisfying

$$\rho_{t} = \rho_{0}\bar{F}(t) + \int_{0}^{t} z(u)\lambda(u)\bar{G}(t-u)du,
z(t)\lambda(t) = \begin{cases} \lambda(t), & \rho_{t} < 1, \\ \rho_{0}f(t) + \int_{0}^{t} z(u)\lambda(u)g(t-u)du, & \rho_{t} = 1, \end{cases}$$
(3.40)

and

$$\mathbb{1}_{\{\rho_t < 1\}} \le z(t) \le 1 \quad a.e. \ in [0, T]. \tag{3.41}$$

is a solution to (3.33)-(3.34). In addition, the function $z\lambda$ is unique almost everywhere.

Proof. Since z(t) is bounded and $\lambda(t), g(t) \in L^1[0,T]$, by Young's convolution inequality we have

$$\frac{\partial}{\partial t}z(u)\bar{G}(t-u)\lambda(u) = -z(u)\lambda(u)g(t-u) \in L^1([0,T] \times [0,T]).$$

Therefore, by [35, Thm 2.7] we have for $t \in [0, T]$

$$\rho'_{t} = -\rho_{0}f(t) + z(t)\lambda(t) - \int_{0}^{t} z(u)g(t-u)\lambda(u)du, \quad \text{a.e. in } [0,T].$$
(3.42)

Recall that $\{\tau_i, \sigma_i\}_{i \in \mathbb{N}}$ are defined in (3.35). For any $i = 1, 2, 3, \dots, t \in (\tau_i, \sigma_i)$, we have $\rho_t = 1$. Consequently $\rho'_t = 0$ in these intervals. By (3.42) we thus have for almost every $t \in (\tau_i, \sigma_i)$

$$-\rho_0 f(t) + z(t)\lambda(t) - \int_0^{\tau_i} \lambda(u)g(t-u)du - \int_{\tau_i}^t z(u)\lambda(u)g(t-u)du = 0.$$
 (3.43)

By [7, Thm 6.3.1] we know that for $t > \tau_i$ there exist a unique solution $x(t) \in L^1_{loc}(\mathbb{R}_+)$ of the Volterra integral equation

$$x(t) = \rho_0 f(t) + \int_0^{\tau_i} \lambda(u) g(t - u) du + \int_{\tau_i}^t x(u) g(t - u) du.$$
 (3.44)

Since by (3.43) we have $z\lambda$ is a solution to (3.44), by the uniqueness of the solution we can conclude that $x(t) = z(t)\lambda(t)$ for $t \in (\tau_i, \sigma_i)$.

Remark 3.2. In the proof of Theorem 3.2 we can see that when $\lambda(t) > 0$, the solution z(t) is unique almost surely.

Remark 3.3. Notice that if, in addition to Assumption 3.2, we assume g > 0 then $\lambda(t) > 0$ a.e. when $\rho_t = 1$. That is, for almost every $t \in [\tau_i, \sigma_i]$ we must have $\lambda(t) > 0$, where $\{\tau_i, \sigma_i\}_{i \in \mathbb{N}}$ are defined in (3.35). This can be proved by contradiction. Assume $\lambda(t) = 0$ for some positive measure set $K \subset [\tau_i, \sigma_i]$. Without loss of generality we can assume $K = (t', t' + \delta) \subset [\tau_i, \sigma_i]$. By (3.33) we have

$$\rho_{t'} = \rho_0 \bar{F}(t') + \int_0^{\tau_i} z(u) \bar{G}(t'-u) \lambda(u) du + \int_{\tau_i}^{t'} z(u) \bar{G}(t'-u) \lambda(u) du, \tag{3.45}$$

and

$$\rho_{t'+\delta} = \rho_0 \bar{F}(t'+\delta) + \int_0^{\tau_i} z(u)\bar{G}(t'+\delta-u)\lambda(u)du + \int_{\tau_i}^{t'} z(u)\bar{G}(t'+\delta-u)\lambda(u)du. \tag{3.46}$$

Since \bar{F} is non-increasing and \bar{G} is strictly decreasing, by (3.45)-(3.46) we have $\rho_{t'+\delta} < \rho_{t'}$. This is a contradiction since $\rho_t = 1$ for $t \in (\tau_i, \sigma_i)$.

Remark 3.4. Note that the proof of Theorem 3.2, provides a characterization for σ_i . Indeed, σ_i equals the first time after τ_i that $x(t) > \lambda(t)$ for a positive measure set. To see that, notice by (3.41), for $t \in (\tau_i, \sigma_i)$ we have $x(t) = z(t)\lambda(t) \le \lambda(t)$. Suppose there exist $\varepsilon > 0$ such that $x(t) \le \lambda(t)$ for almost every $t \in [\sigma_i, \sigma_i + \varepsilon)$, then choose \tilde{z} such that \tilde{z} satisfies (3.41) and $\tilde{z}(t)\lambda(t) = x(t)$ for $t \in [\sigma_i, \sigma_i + \varepsilon)$ and, $\tilde{z}(t) = z(t)$ for $t \in [0, \sigma_i)$. By (3.42) and (3.44) we know that there exists a function $\tilde{\rho}$ such that

$$\tilde{\rho}_t = \begin{cases} \rho_t & t \in [0, \sigma_i) \\ \rho_0 \bar{F}(t) + \int_0^t \tilde{z}(u) \lambda(u) \bar{G}(t-u) du & t \in [\sigma_i, \sigma_i + \varepsilon) \end{cases},$$

and $\tilde{\rho}_t' = 0$ for $t \in [\sigma_i, \sigma_i + \varepsilon)$. This implies that $\tilde{\rho}_t = 1$ in this interval. However, $\rho_t < 1$ for $t \in [\sigma_i, \sigma_i + \varepsilon)$. From the uniqueness of ρ established by Theorem 3.1 we can concluded that this is a contradiction.

Since we have obtained the unique solvability of ρ , we can establish the fluid limit result for the entire sequence \bar{N}^n .

Theorem 3.3. Let Assumption 3.2 hold. Then

(i) For any T > 0, there exists a real-valued continuous deterministic process ρ such that almost surely,

$$\lim_{n \to \infty} \sup_{t \in [0,T]} |\bar{N}_t^n - \rho_t| = 0. \tag{3.47}$$

(ii) Moreover, there exists a bounded function w such that almost surely

$$\mathbb{1}_{\{\bar{N}_{t}^{n} < 1\}} \stackrel{*}{\rightharpoonup} w(t) \quad in \ L^{\infty}[0, T], \tag{3.48}$$

 λ -almost surely in t, where w solves (3.40)-(3.41).

(iii) Furthermore, ρ and w defined in (3.47)-(3.48) satisfy

$$\rho_t = \rho_0 \bar{F}(t) + \int_0^t w(u) \bar{G}(t - u) \lambda(u) du, \quad t \in [0, T], \quad and$$

$$\mathbb{1}_{\{\rho_u - < 1\}} \le w(u) \le 1, \quad a.e. \text{ in } [0, T].$$
(3.49)

That is, (ρ, w) as in (3.49) is a solution, interpreted according to Definition 2.2, to the following non-linear discontinuous Volterra integral equation

$$\rho_t = \rho_0 \bar{F}(t) + \int_0^t \mathbb{1}_{\{\rho_{u-} < 1\}} \bar{G}(t-u) \lambda(u) du.$$
 (3.50)

Proof. Part (i). From Proposition 3.2, for any subsequence there exists a further subsequence (r_k) such that almost surely

$$\bar{N}_{t}^{r_{k}} \rightarrow \rho_{t}$$

in the uniform topology, where ρ solves (3.50) path by path. By Theorem 3.1, ρ is unique. Consequently ρ is a deterministic function. Moreover, from the uniqueness of ρ again we can conclude that the entire sequence \bar{N}^n converges to ρ almost surely in uniform topology.

Part (ii). By Proposition 3.2 we have for every subsequence there exists a subsubsequence (r_k) and a bounded, possibly stochastic process w such that almost surely

$$\mathbb{1}_{\{\bar{N}^{r_k} < 1\}} \lambda(u) \stackrel{*}{\rightharpoonup} w(u)\lambda(u) \quad \text{in } L^{\infty}[0, T]. \tag{3.51}$$

By Theorem 3.2 we know that this $w\lambda$ is unique. Therefore the weak-star convergence in (3.51) holds for the entire sequence.

Part (iii). This follows directly from Proposition 3.2.(iii) and Theorems 3.1-3.2.
$$\Box$$

Note that the probability that an incoming arrival at time t will be accepted to the system is given by $P(\rho_{t-}^n < 1)$. The following corollary provides asymptotics for this acceptance probability.

Corollary 3.1. The acceptance probability in the n-th $M_t/G/n/n$ model $\mathbb{P}(\bar{N}_{t-}^n < 1)$ satisfies the following convergence

$$\mathbb{P}\left(\bar{N}_{u-}^{n}<1\right)\to w(u),\quad for\ \lambda\text{-almost every }u\in[0,T],$$

where w is defined in Theorem 3.3.

Proof. $\mathbb{1}_{\{\bar{N}_{u-}^n<1\}}$ is piecewise constant with almost surely finitely many jumps in [0,T] since the number of jumps of \bar{N}^n is bounded by twice that of the arrivals. By Theorem 2.1 we obtain the tightness of $(\mathbb{1}_{\{\bar{N}_{u-}^n<1\}})$. By (3.48) we have λ -almost surely, for any $\phi \in L^1[0,T]$

$$\lim_{k \to \infty} \int_0^t \phi(u) \mathbb{1}_{\{\bar{N}_{u-}^n < 1\}} du = \int_0^t \phi(u) w(u) du.$$

By taking $\phi(\cdot) = \delta_{(\cdot)}$ we can conclude that the finite dimensional distributions of $\mathbb{1}_{\{\bar{N}_{t-}^n < 1\}}$ converge to that of w(t). By Theorem 2.2 we have $\mathbb{1}_{\{\bar{N}_{n-}^n < 1\}} \Rightarrow w(u)$. This implies that $\mathbb{1}_{\{\bar{N}_{n-}^n < 1\}} \lambda(u) \Rightarrow$

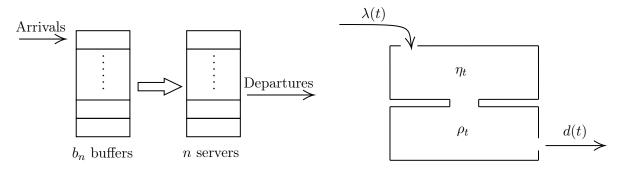


FIGURE 2. Loss system with buffer and its fluid model

 $w(u)\lambda(u)$. By Theorem 3.2 we have $w\lambda$ is unique and thus deterministic. Therefore, the convergence becomes

$$\mathbb{1}_{\{\bar{N}^n_{::} < 1\}} \lambda(u) \xrightarrow{p} w(u) \lambda(u),$$

in the Skorokhod topology. Moreover, since the indicator functions are uniformly bounded and λ is integrable in [0,T], $(\mathbb{1}_{\{\bar{N}_{u-}^n<1\}}\lambda(u))$ are uniformly integrable. By [10, Thm 5.5.2] we have for almost every $u \in [0,T]$,

$$\mathbb{E}\left[\mathbb{1}_{\{\bar{N}_{u-}^n<1\}}\lambda(u)\right] = \mathbb{P}\left(\bar{N}_{u-}^n<1\right)\lambda(u) \to w(u)\lambda(u).$$

Remark 3.5. As noticed in (3.40), the function w(t) can be discontinuous at τ_i even when λ is continuous. This means the limit of the blocking/acceptance probability is discontinuous. This property is further reflected in the numerics below.

4. Fluid Limit for Loss System with Buffer

4.1. **Setup.** In this section, we introduce a time-varying many-server loss queuing model with buffer. We work with a sequence of queuing systems indexed by n, subject to the following assumptions.

Assumption 4.1. We consider a $M_t/G/n/n + b_n$ loss queuing system; namely, a queuing model with

- i. a nonhomogeneous Poisson arrival process A^n with rate or intensity function $n\lambda(\cdot)$, where λ is locally integrable;
- ii. general customer service times sampled independently from a distribution G with density g bounded by a constant $c_g > 0$;
- iii. the system has n servers and b_n buffer spaces or waiting spaces. When the system is full, new incoming customer arrivals are lost. Additionally, b_n satisfies

$$\lim_{n \to \infty} \frac{b_n}{n} \to \beta. \tag{4.1}$$

4.2. Characterization of Relevant Stochastic Processes. Let S_t^n and Q_t^n denote the number of customers in service and buffer, respectively, at time t. In addition, let D_t^n denote the cumulative number of departures from the system by time t. For the scaled processes, we define

$$\bar{S}_t^n := \frac{S_t^n}{n}, \ \bar{Q}_t^n := \frac{Q_t^n}{n}, \ \text{and} \ \bar{D}_t^n := \frac{D_t^n}{n},$$

to be the n-scaled number in service, in buffer and of departures respectively. Also, let \mathcal{F}_t^n be the filtration generated by $\{\bar{S}_s^n, \bar{Q}_s^n : s \in [0, t]\}$. Let T_i, V_i , and D_i represent respectively the arrival

time, service time, and departure time of the *i*-th customer to the system. Note that a customer who arrives to find at least one idle server has their arrival time coincide with their service start time. However, a customer who upon arrival finds all servers busy and must first enter the buffer to wait, has their service start time determined by the arrival and service times of prior customers. In addition, their service entry time coincides with the departure time of a prior customer. For this scenario, we let V_{j_i} denote the service time of the customer who enters service at time D_i . For simplicity, we initially assume that the number of customers in the system at time t = 0 is zero. This assumption will be relaxed in the sequel.

- 4.2.1. Busy servers or customers in service. Observe that the number of busy servers or the number in service at time t consists of customers from two groups:
- (a) Customers admitted directly upon arrival. This scenario is similar to the setup of Section 3. Observe that the number of customers at time t, who were directly admitted upon arrival, consists of all arrivals to the system such that all of the following conditions are met:
- (i) the customer arrival occurs at or prior to time t,
- (ii) the number of occupied servers upon the customer's arrival is less than n, and
- (iii) the remaining service time of this customer at time t is positive, that is, the customer is yet to depart the system.

For the *i*-th customer arriving to the system, these conditions correspond to $\{T_i \leq t\}$, $\{S_{T_i}^n < n\}$ or $\{\bar{S}_{T_i}^n < 1\}$, and $\{V_i > t - T_i\}$ respectively. Consequently, the number of customers at time t, who were directly admitted upon arrival equals:

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{S_{T_i}^n < n\}} \mathbb{1}_{\{V_i > t - T_i\}}.$$
(4.2)

- (b) Customers promoted from the buffer. The customers in this scenario start service at the departure time D_i of some customer i. Observe that the number of customers at time t, who were promoted from the buffers, consists of all departures such that all of the following conditions are met:
- (i) the service start time D_i of this customer is at or prior to time t,
- (ii) the buffer is non-empty at time D_i , and
- (iii) the remaining service time of this customer at time t is positive, that is, the customer is yet to depart the system.

For the customer promoted from buffer at time D_i , these conditions correspond to $\{D_i \leq t\}$, $\{Q_{D_{i-}}^n > 0\}$ or $\{\bar{Q}_{D_{i-}}^n > 0\}$, and $\{V_{j_i} > t - D_i\}$ respectively. Consequently, the number of customers at time t, who were promoted from the buffer equals:

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\{Q_{D_i}^n > 0\}} \mathbb{1}_{\{D_i + V_{j_i} > t\}}.$$
(4.3)

Therefore, by combining the two groups of customers from (4.2)-(4.3), we have the number of customers in service at time t satisfies

$$S_t^n = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{S_{T_i}^n < n\}} \mathbb{1}_{\{V_i > t - T_i\}} + \sum_{i=1}^{\infty} \mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\{Q_{D_i}^n > 0\}} \mathbb{1}_{\{D_i + V_{j_i} > t\}}. \tag{4.4}$$

On scaling (4.4) by n, we have in contrast to (3.2) that the scaled number of busy servers satisfy

$$\bar{S}_{t}^{n} = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{T_{i} \leq t\}} \mathbb{1}_{\{\bar{S}_{T_{i}}^{n} < 1\}} \mathbb{1}_{\{V_{i} > t - T_{i}\}} + \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{D_{i} \leq t\}} \mathbb{1}_{\{\bar{Q}_{D_{i}}^{n} > 0\}} \mathbb{1}_{\{D_{i} + V_{j_{i}} > t\}}. \tag{4.5}$$

- 4.2.2. Occupied buffers. The number of occupied buffers equals the difference between two groups:
- (a) Customers that entered the buffer. Observe that the total number of customers who entered the buffer by time t consists of those individuals who satisfy all of the following conditions:
- (i) the customer arrival occurs at or prior to time t,
- (ii) the number of occupied servers upon the customer's arrival is n, and
- (iii) the buffer upon the customer's arrival is not full.

For the i-th customer arriving to the system, these conditions correspond to $\{T_i \leq t\}$, $\{S^n_{T_{i-}} = n\}$ or $\{\bar{S}^n_{T_{i-}} = 1\}$, and $\{Q^n_{T_{i-}} < b_n\}$ or $\{\bar{Q}^n_{T_{i-}} < \frac{b_n}{n}\}$ respectively. Consequently, the number of customers who entered the buffer by time t equals:

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{S_{T_i}^n = n\}} \mathbb{1}_{\{Q_{T_i}^n < b_n\}}. \tag{4.6}$$

- (b) Customers that exited the buffer. The customers in this scenario start service at the departure time D_i of some customer i. Observe that the total number of customers who departed from the buffer by time t consists of those individuals who satisfy all of the following conditions:
- (i) the service start time D_i of this customer is at or prior to time t, and
- (ii) the buffer is non-empty at time D_i .

For the customer departing from buffer at time D_i , these conditions correspond to $\{D_i \leq t\}$ and $\{Q_{D_i-}^n > 0\}$ or $\{\bar{Q}_{D_i-}^n > 0\}$ respectively. Consequently, the number of customers that exited the buffer by time t equals:

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\{Q_{D_i}^n > 0\}}. \tag{4.7}$$

Therefore, by taking the difference between the two groups of customers from (4.6)-(4.7), we have the number of customers in buffer at time t satisfies

$$Q_t^n = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{S_{T_i}^n = n\}} \mathbb{1}_{\{Q_{T_i}^n < b_n\}} - \sum_{i=1}^{\infty} \mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\{Q_{D_i}^n > 0\}}. \tag{4.8}$$

On scaling (4.8) by n, this yields that the scaled number in buffer satisfy

$$\bar{Q}_{t}^{n} = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{T_{i} \leq t\}} \mathbb{1}_{\{\bar{S}_{T_{i}}^{n} = 1\}} \mathbb{1}_{\{\bar{Q}_{T_{i}}^{n} < \frac{b_{n}}{n}\}} - \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{D_{i} \leq t\}} \mathbb{1}_{\{\bar{Q}_{D_{i}}^{n} > 0\}}. \tag{4.9}$$

- 4.2.3. Departures. Observe that the cumulative number of departures also include customers from two groups:
- (a) Departure of customers admitted directly upon arrival. Observe that the number of customers at time t, who were directly admitted upon arrival and then departed from the system, consists of those customers who satisfy all of the following conditions:
- (i) the customer arrival occurs at or prior to time t,
- (ii) the number of occupied servers upon the customer's arrival is less than n, and
- (iii) the customer has departed from the system by time t.

For the *i*-th customer arriving to the system, these conditions correspond to $\{T_i \leq t\}$, $\{S_{T_i}^n < n\}$ or $\{\bar{S}_{T_i}^n < 1\}$, and $\{T_i + V_i \leq t\}$ respectively. Consequently, the number of customers at time t, who were admitted directly upon arrival and then departed equals:

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{S_{T_i}^n < n\}} \mathbb{1}_{\{T_i + V_i \le t\}}. \tag{4.10}$$

(b) Departure of customers promoted from the buffer. The customers in this scenario start service at the departure time D_i of some customer i. Observe that the number of customers at time t, who

were promoted from the buffers and then departed from the system, consists of those customers who satisfy all of the following conditions:

- (i) the service start time D_i of this customer is at or prior to time t,
- (ii) the buffer is non-empty at time D_i , and
- (iii) the customer has departed from the system at time t.

For the customer promoted from buffer at time D_i , these conditions correspond to $\{D_i \leq t\}$, $\{Q_{D_{i-}}^n > 0\}$ or $\{\bar{Q}_{D_{i-}}^n > 0\}$, and $\{D_i + V_{j_i} \leq t\}$ respectively. Consequently, the number of customers at time t, who were promoted from the buffer and departed equals:

$$\sum_{i=1}^{\infty} \mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\{Q_{D_i}^n > 0\}} \mathbb{1}_{\{D_i + V_{j_i} \le t\}}. \tag{4.11}$$

Therefore, by combining the two groups of customers from (4.10)-(4.11) we have the cumulative departures at time t satisfies

$$D_t^n = \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \le t\}} \mathbb{1}_{\{S_{T_i}^n < n\}} \mathbb{1}_{\{V_i + T_i \le t\}} + \sum_{i=1}^{\infty} \mathbb{1}_{\{Q_{D_i}^n > 0\}} \mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\{D_i + V_{j_i} \le t\}}.$$

On scaling by n we have

$$\bar{D}_{t}^{n} = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{D_{i} \leq t\}} = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{T_{i} \leq t\}} \mathbb{1}_{\{\bar{S}_{T_{i}}^{n} < 1\}} \mathbb{1}_{\{V_{i} + T_{i} \leq t\}} + \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{\bar{Q}_{D_{i}}^{n} > 0\}} \mathbb{1}_{\{D_{i} + V_{j_{i}} \leq t\}}.$$
(4.12)

4.3. **Stochastic Integral Representation.** As in Section 3, we will use random measures to obtain cleaner representations of the processes under consideration. To that effect, we define:

$$\begin{split} W_{n}^{s,A}(t,u,x) &= \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{u \leq t\}} \mathbb{1}_{\bar{S}_{u-}^{n} < 1\}} \mathbb{1}_{\{x > t - u\}}, \qquad W_{n}^{s,D}(t,u,x) = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{u \leq t\}} \mathbb{1}_{\bar{Q}_{u-}^{n} > 0\}} \mathbb{1}_{\{x > t - u\}}, \\ W_{n}^{q,A}(t,u,x) &= \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{u \leq t\}} \mathbb{1}_{\bar{S}_{u-}^{n} = 1\}} \mathbb{1}_{\bar{Q}_{u-}^{n} < \frac{b_{n}}{n}}, \qquad W_{n}^{q,D}(t,u,x) = \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{u \leq t\}} \mathbb{1}_{\bar{Q}_{u-}^{n} > 0\}}, \\ W_{n}^{d,A}(t,u,x) &= \frac{1}{n} \sum_{i=1}^{\infty} \mathbb{1}_{\{u \leq t\}} \mathbb{1}_{\bar{Q}_{u-}^{n} > 0\}} \mathbb{1}_{\{x \leq t - u\}}. \end{split}$$

Using these notations, the relations (4.5), (4.9) and (4.12) can be expressed as stochastic integrals

$$\bar{S}_{t}^{n} = \int_{0}^{t} \int_{\mathbb{R}} W_{n}^{s,A}(t,u,x) \mathcal{M}^{n,A}(du,dx) + \int_{0}^{t} \int_{\mathbb{R}} W_{n}^{s,D}(t,u,x) \mathcal{M}^{n,D}(du,dx), \tag{4.13}$$

$$\bar{Q}_{t}^{n} = \int_{0}^{t} \int_{\mathbb{R}} W_{n}^{q,A}(t,u,x) \mathcal{M}^{n,A}(du,dx) - \int_{0}^{t} \int_{\mathbb{R}} W_{n}^{q,D}(t,u,x) \mathcal{M}^{n,D}(du,dx), \tag{4.14}$$

$$\bar{D}_{t}^{n} = \int_{0}^{t} \int_{\mathbb{R}} W_{n}^{d,A}(t,u,x) \mathcal{M}^{n,A}(du,dx) + \int_{0}^{t} \int_{\mathbb{R}} W_{n}^{d,D}(t,u,x) \mathcal{M}^{n,D}(du,dx), \tag{4.15}$$

where $\mathcal{M}^{n,A}$ is the counting measure associated with the marked point process of the arrival and service time pairs (T_i, V_i) , and $\mathcal{M}^{n,D}$ is the counting measure associated with the marked point process of the departure and service time pairs (D_i, V_{j_i}) . Since the number of cumulative departures in [0, t] is bounded by the number of arrivals in the same interval, the departure process is a locally finite point process. Recall Definition 2.1 and denote the intensity measure of the scaled departure process of the n-th model to be ν_n . The following proposition shows that the scaled departure process exhibits an intensity or rate function.

Proposition 4.1. Let Assumption 4.1 hold. Then,

(i) For every $n \in \mathbb{N}$, the intensity measure of the scaled departure process for the n-th model, ν_n is absolutely continuous w.r.t. Lebesgue measure. That is, there exists a density function d_n for every ν_n such that

$$\mathbb{E}[\bar{D}_t^n] = \nu_n(0, t] = \int_0^t d_n(u) du.$$

(ii) There exists a bounded function d on [0,T] and a subsequence (n_k) such that

$$\lim_{k \to \infty} \sup_{t \in [0,T]} \left| \mathbb{E}[\bar{D}_t^{n_k}] - D_t \right| = 0,$$

where $D_t = \int_0^t d(u)du$.

(iii) Furthermore,

$$d_{n_k} \stackrel{*}{\rightharpoonup} d$$
 in $L^{\infty}[0,T]$.

Proof. Part (i). For any n, denote the service start time of the k-th customer to be T'_k . Define the departure process of the k-th customer from the i-th server by $D_t^{k,i,n}$ and the corresponding occupancy indicator of the *i*-th server $\boldsymbol{B}_t^{k,i,n}$ as following:

$$D_t^{k,i,n} = \mathbb{1}_{\{T_k' + V_k \le t\}}, \quad B_t^{k,i,n} = \mathbb{1}_{\{T_k' \le t < T_k' + V_k\}}.$$

Define the hazard rate

$$h(x) := \frac{g(x)}{1 - G(x)}, \quad x \in [0, M) \text{ where } M := \sup\{x \in [0, \infty) : G(x) < 1\}.$$

Note that h(u) is almost surely well-defined on $[0, V_k]$. Let $\mathcal{F}_t^{k,i} := \sigma\{B_s^{k,i,n}, \text{ for } 0 \leq s < t\}$. We claim that the process

$$X_t^{k,i,n} = D_t^{k,i,n} - \int_0^t B_u^{k,i,n} h(u - T_k') du, \quad t \ge 0,$$
(4.16)

is a martingale w.r.t. $\mathcal{F}_t^{k,i}$. It suffices to consider the following elements of $\mathcal{F}_s^{k,i}$ for $0 \le s < t$:

- $\begin{array}{ll} \text{(a)} \ \{T_k' = r, V_k = v\} \ \text{for} \ r + v \leq s, \\ \text{(b)} \ \{T_k' = r, T_k' + V_k > s\} \ \text{for} \ r \leq s, \ \text{and} \\ \text{(c)} \ \{T_k' > s\}. \end{array}$
- (a) For $r + v \le s$, we have

$$\mathbb{E}\left[X_{t}^{k,i,n} \mid T_{k}' = r, V_{k} = v\right] \\
= \mathbb{P}\left(T_{k}' + V_{k} \le t \mid T_{k}' = r, V_{k} = v\right) - \int_{r}^{t} h(u - r)\mathbb{P}\left(u < T_{k}' + V_{k} \mid T_{k}' = r, V_{k} = v\right) du \\
= 1 - \int_{r}^{r+v} h(u - r) du, \tag{4.17}$$

where the last expression is the value of $X_s^{k,i,n}$ on $\{T_k'=r,V_k=v\}$.

(b) For $r \leq s$ we have

$$\mathbb{E}\left[X_{t}^{k,i,n} \mid T_{k}' = r, T_{k}' + V_{k} > s\right] \\
= \mathbb{P}\left(T_{k}' + V_{k} \leq t \mid T_{k}' = r, T_{k}' + V_{k} > s\right) - \int_{r}^{t} h(u - r) \mathbb{P}\left(u < T_{k}' + V_{k} \mid T_{k}' = r, T_{k}' + V_{k} > s\right) du \\
= \frac{\mathbb{P}\left(s < T_{k}' + V_{k} \leq t \mid T_{k}' = r\right)}{\mathbb{P}\left(T_{k}' + V_{k} > s \mid T_{k}' = r\right)} - \int_{r}^{t} h(u - r) \frac{\mathbb{P}\left(T_{k}' + V_{k} > u \vee s \mid T_{k}' = r\right)}{\mathbb{P}\left(T_{k}' + V_{k} > s \mid T_{k}' = r\right)} du \\
= \frac{G(t - r) - G(s - r)}{1 - G(s - r)} - \int_{r}^{t} h(u - r) \frac{1 - G(u \vee s - r)}{1 - G(s - r)} du, \tag{4.18}$$

where using the definition of h in the last integral

$$\int_{r}^{t} h(u-r) \frac{1 - G(u \vee s - r)}{1 - G(s - r)} du = \int_{r}^{s} h(u - r) du + \int_{s}^{t} \frac{g(u - r)}{1 - G(s - r)} du$$

$$= \int_{r}^{s} h(u - r) du + \frac{G(t - r) - G(s - r)}{1 - G(s - r)}. \quad (4.19)$$

Therefore, plugging (4.19) into (4.18) we have

$$\mathbb{E}\left[X_t^{k,i,n} \mid T_k' = r, T_k' + V_k > s\right] = -\int_r^s h(u - r) du,\tag{4.20}$$

where the right hand side is the value of $X_s^{k,i,n}$ on $\{T_k'=r,T_k'+V_k>s\}$.

(c) Finally, consider

$$\mathbb{E}\left[X_t^{k,i,n} \mid T_k' > s\right] = \mathbb{P}\left(T_k' + V_k \le t \mid T_k' > s\right) - \mathbb{E}\left[\int_0^t h(u - T_k') \mathbb{1}\left(T_k' \le u < T_k' + V_k\right) du \mid T_k' > s\right] \\
= \frac{1}{\mathbb{P}\left(T_k' > s\right)} \left(\mathbb{P}\left(T_k' + V_k \le t, T_k' > s\right) - \mathbb{E}\left[\mathbb{1}_{\{T_k' > s\}} \int_{T_k'}^{(T_k' + V_k) \wedge t} h(u - T_k') du\right]\right). \tag{4.21}$$

The last term of the numerator in (4.21) can be expressed as

$$\mathbb{E}\left[\mathbb{1}_{\{T'_k>s\}}\int_{T'_k}^{(T'_k+V_k)\wedge t}h(u-T'_k)du\right] = \mathbb{E}\left[\mathbb{1}_{\{T'_k>s\}}\mathbb{E}\left[\int_{T'_k}^{(T'_k+V_k)\wedge t}h(u-T'_k)du\,\middle|\,T'_k\right]\right] \\
= \mathbb{E}\left[\mathbb{1}_{\{T'_k>s\}}\mathbb{E}\left[\int_0^{(T'_k+V_k)\wedge t-T'_k}h(u)du\,\middle|\,T'_k\right]\right], \qquad (4.22)$$

where elementary integration yields:

$$\mathbb{E}\left[\int_{0}^{\left(T'_{k}+V_{k}\right)\wedge t-T'_{k}}h(u)du\,\middle|\,T'_{k}=r\right]=\mathbb{E}\left[-\log\left\{1-G\left(\left(T'_{k}+V_{k}\right)\wedge t-T'_{k}\right)\right\}\,\middle|\,T'_{k}=r\right]$$

$$=G(t-r). \quad (4.23)$$

Plugging (4.23) into (4.22) we have

$$\mathbb{E}\left[\mathbb{1}_{\{T'_k > s\}} \int_{T'_k}^{(T'_k + V_k) \wedge t} h(u - T'_k) du\right] = \mathbb{E}\left[G(t - T'_k)\mathbb{1}_{\{T'_k > s\}}\right] = \mathbb{P}\left(T'_k + V_k \le t, T'_k > s\right). \quad (4.24)$$

Using (4.24) in (4.21) we obtain

$$\mathbb{E}\left[X_t^{k,i,n} \mid T_k' > s\right] = 0,\tag{4.25}$$

which is exactly the value of $X_s^{k,i,n}$ on $\{T_k' > s\}$.

Combining our conclusions from cases (a)-(c) given by relations (4.17), (4.20) and (4.25) we can conclude that

$$\mathbb{E}\left[X_t^{k,i,n} \mid \mathcal{F}_s\right] = X_s^{k,i,n}.$$

This proves our claim that $X_t^{k,i,n}$ given by (4.16) is a martingale w.r.t. $\mathcal{F}_t^{k,i}$. Consequently we have

$$\nu_n(0,t] = \mathbb{E}\left[\bar{D}_t^n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i,k}D_t^{k,i,n}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i,k}\int_0^t B_u^{k,i,n}h(u-T_k')du\right]. \tag{4.26}$$

Since our integrands are non-negative, by Tonelli's theorem, we interchange expectation and integral to obtain

$$\nu_n(0,t] = \int_0^t \mathbb{E}\left[\frac{1}{n}\sum_{i,k} B_u^{k,i,n} h(u - T_k')\right] du.$$
 (4.27)

This implies that ν_n is absolutely continuous w.r.t. Lebesgue measure. Denoted

$$d_n(u) = \mathbb{E}\left[\frac{1}{n}\sum_{i,k}B_u^{k,i,n}h(u-T_k')\right],$$

be the intensity function in (4.27).

Part (ii). Since d_n are non-negative, and

$$\int_0^t d_n(u)du = \mathbb{E}\left[\bar{D}_t^n\right] \le \int_0^T \lambda(u)du < \infty,\tag{4.28}$$

are uniformly bounded, by Helly's selection theorem there exists a bounded non-decreasing function D and a subsequence (n_k) such that the pointwise convergence $\int_0^t dn_k(u)du \to D_t$ holds. Furthermore, since D is continuous, then the convergence is uniform (see for example [34, Sec 0.1]). It remains to show that D_t is absolute continuous with a non-negative density d. Since $\sum_i B_t^{k,i,n} = \mathbb{1}_{\{T_k' \le t < T_k' + V_k\}}$, from (4.26) we have for any n

$$\nu_n(s,t] = \mathbb{E}\left[\frac{1}{n}\sum_{i,k}\int_s^t B_u^{k,i,n}h(u-T_k')du\right] = \mathbb{E}\left[\frac{1}{n}\sum_k\int_s^t \mathbb{1}_{\{T_k'\leq u< T_k'+V_k\}}h(u-T_k')du\right].$$

Therefore,

$$\begin{split} \nu_n(s,t] &= \frac{1}{n} \mathbb{E}\left[\sum_k \int_{T_k' \vee s}^{\left(T_k' + V_k\right) \wedge t} h(u - T_k') du\right] = \frac{1}{n} \mathbb{E}\left[\sum_k \mathbb{E}\left[\int_{\left(T_k' \vee s\right) - T_k'}^{\left(T_k' + V_k\right) \wedge t - T_k'} h(u) du \mid T_k'\right]\right] \\ &= \frac{1}{n} \mathbb{E}\left[\sum_k \mathbb{E}\left[\log\left(1 - G\left(\left(s - T_k'\right) \vee 0\right)\right) - \log\left(1 - G\left(\left(T_k' + V_k\right) \wedge t - T_k'\right)\right) \mid T_k'\right]\right]. \end{split}$$

Using (4.23) the above equation becomes

$$\nu_n(s,t] = \frac{1}{n} \mathbb{E}\left[\sum_k \log\left(1 - G\left(\left(s - T_k'\right) \vee 0\right)\right) + G(t - T_k')\right]. \tag{4.29}$$

Denote $A_k = \log(1 - G((s - T_k') \vee 0)) + G(t - T_k')$. Recall in Assumption 4.1 that $g(x) \leq c_g$. For $s < T_k'$ we have

$$A_k = \log(1 - G(0)) + G(t - T_k') \le G(t - T_k') \le G(t - s) \le c_g(t - s), \tag{4.30}$$

where the last inequality follows from mean value theorem. On the other hand, since $\log(x) \le x - 1$, for $s \ge T'_k$ we have

$$A_k = \log(1 - G(s - T_k')) + G(t - T_k') \le -G(s - T_k') + G(t - T_k') \le c_g(t - s). \tag{4.31}$$

Combining (4.29)-(4.31) we obtain

$$\nu_n(s,t] \le \frac{1}{n} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}_{\{T'_k \le t\}} \right] c_g(t-s),$$

Since $\mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}_{\{T_k' \le t\}}] \le \mathbb{E}[A_t^n] \le \int_0^t n\lambda(u)du$, we have for any n,

$$\nu_n(s,t] \le c_g(t-s) \int_0^T \lambda(u) du. \tag{4.32}$$

Therefore

$$D_t - D_s \le c_g(t - s) \int_0^T \lambda(u) du. \tag{4.33}$$

By (4.33), for $\varepsilon > 0$ there exists $\delta = \varepsilon/(c_g \int_0^T \lambda(u) du)$ such that for any finite set of disjoint intervals $(a_1, b_1), \ldots, (a_K, b_K)$ satisfying $\sum_{j=1}^K (b_j - a_j) < \delta$,

$$\sum_{j=1}^{K} \left| D_{b_j} - D_{a_j} \right| \le c_g \int_0^T \lambda(u) du \sum_{j=1}^{K} (b_j - a_j) < \varepsilon.$$

By [13, Prop 3.32], we conclude that D is absolutely continuous w.r.t. Lebesgue measure. By Radon–Nikodym theorem there exists a density function d such that $D_t = \int_0^t d(u)du$. Finally, notice from (4.32)-(4.33) we know that d_{n_k} and d are bounded by $c_g \int_0^T \lambda(u)du$. This completes the proof of Part (ii).

Part (iii). The convergence $\int_0^t d_{n_k}(u)du \to D_t$ implies that for any $0 \le s < t \le T$ we have

$$\lim_{k \to \infty} \int_s^t d_{n_k}(u) du = \lim_{k \to \infty} \int_0^t \mathbb{1}_{[s,t)} d_{n_k}(u) du = \int_0^t \mathbb{1}_{[s,t)} d(u) du.$$

This can be extended to any step function q to give

$$\lim_{k \to \infty} \int_0^t q(u) dn_k(u) du = \int_0^t q(u) d(u) du.$$

Since step functions are dense in L^1 , we can conclude that for any $\phi \in L^1[0,T]$

$$\lim_{k \to \infty} \int_0^t \phi(u) dn_k(u) du = \int_0^t \phi(u) d(u) du.$$

Thanks to the existence of the density functions d_n from Proposition 4.1, we can characterize the intensities of the random measures under consideration.

Lemma 4.1. Let Assumption 4.1 hold. The intensity measures of the marked point processes $\mathcal{M}^{n,A}$ and $\mathcal{M}^{n,D}$ are

$$\mathcal{M}_{c}^{n,A}(du,dx) = \mathbb{E}\left[\mathcal{M}^{n,A}(du,dx)\right] = n\lambda(u)g(x)dudx,$$

$$\mathcal{M}_{c}^{n,D}(du,dx) = \mathbb{E}\left[\mathcal{M}^{n,D}(du,dx)\right] = nd_{n}(u)g(x)dudx.$$
(4.34)

Proof. The first part is trivial and has already been utilized in Section 3. We show only for $\mathcal{M}^{n,D}$ here. Observe that

$$\mathcal{M}^{n,D}(C \times L) = \sum_{i=1}^{\infty} \mathbb{1}_{C}(D_{i})\mathbb{1}_{L}(V_{j_{i}}),$$

and hence

$$\mathcal{M}_{c}^{n,D}(C \times L) = \mathbb{E}\left[\mathcal{M}^{n,D}(C \times L)\right] = \sum_{i}^{\infty} \mathbb{E}\left[\mathbb{1}_{C}(D_{i})\mathbb{1}_{L}(V_{j_{i}})\right].$$

Note that D_i and V_{i} are independent. Thus we have

$$\mathcal{M}_{c}^{n,D}(C \times L) = \sum_{i=1}^{\infty} \mathbb{E}\left[\mathbb{1}_{C}(D_{i})\right] \mathbb{E}\left[\mathbb{1}_{L}(V_{j_{i}})\right] = \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{C}(D_{i})\right] \mathbb{P}(V_{j_{1}} \in L) = n\nu_{n}(C) \int_{L} g(x) dx. \tag{4.35}$$

Applying Proposition 4.1 to (4.35) we obtain that

$$\mathcal{M}_{c}^{n,D}(C \times L) = \int_{C \times L} n d_{n}(u) g(x) du dx,$$

which proves our desired result.

We now exploit the intensities obtained in Lemma 4.1 to obtain the limit of the stochastic processes $(\bar{S}^n, \bar{Q}^n, \bar{D}^n)$ as n goes to infinity. We again begin with a result proving convergence along a subsequence.

Proposition 4.2. Let Assumption 4.1 hold. Assume that the system starts empty, that is, the number of customers at time 0 is zero. Then

(i) For any T > 0 and for any subsequence, there exists a further subsequence (r_k) and continuous, possibly stochastic processes ρ, η, D such that almost surely

$$\bar{S}_t^{r_k} \to \rho_t, \quad \bar{Q}_t^{r_k} \to \eta_t, \quad \bar{D}_t^{r_k} \to D_t,$$
 (4.36)

in the uniform topology.

(ii) Moreover, given (r_k) , almost surely there exist bounded, possibly stochastic processes w^1, w^2, w^3 such that

$$\mathbb{1}_{\{\bar{S}_{t-}^{r_k}<1\}} \stackrel{*}{\rightharpoonup} w^1(t), \quad \mathbb{1}_{\{\bar{Q}_{t-}^{r_k}>0\}} \stackrel{*}{\rightharpoonup} w^2(t), \quad \mathbb{1}_{\{\bar{Q}_{t-}^{r_k}<\frac{b_n}{n}\}} \stackrel{*}{\rightharpoonup} w^3(t), \quad in \ L^{\infty}[0,T]. \tag{4.37}$$

(iii) Furthermore, almost surely, $(\rho, \eta, D, w^1, w^2, w^3)$ defined in (4.36)-(4.37) satisfy

$$\rho_t = \int_0^t w^1(u)\bar{G}(t-u)\lambda(u)du + \int_0^t w^2(u)\bar{G}(t-u)d(u)du, \tag{4.38}$$

$$\eta_t = \int_0^t (1 - w^1(u))w^3(u)\lambda(u)du - \int_0^t w^2(u)d(u)du, \tag{4.39}$$

$$D_{t} = \int_{0}^{t} w^{1}(u)G(t-u)\lambda(u)du + \int_{0}^{t} w^{2}(u)G(t-u)d(u)du, \tag{4.40}$$

and for almost every $t \in [0, T]$

$$\mathbb{1}_{\{\rho_t < 1\}} \le w^1(t) \le 1, \quad \mathbb{1}_{\{\eta_t > 0\}} \le w^2(t) \le 1, \quad \mathbb{1}_{\{\eta_t < \beta\}} \le w^3(t) \le 1.$$

That is, for almost all $\omega \in \Omega$, $(\rho(\omega), \eta(\omega), D, w^1(\omega), w^2(\omega), w^3(\omega))$ as in (4.38)-(4.40) is a solution, interpreted according to Definition 2.2, to the following non-linear discontinuous Volterra integral

equation

$$\rho_{t} = \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}} \bar{G}(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}} \bar{G}(t-u)d(u)du,
\eta_{t} = \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}=1\}} \mathbb{1}_{\{\eta_{u-}<\beta\}} \lambda(u)du - \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}} d(u)du,
D_{t} = \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}} G(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}} G(t-u)d(u)du.$$
(4.41)

Proof. For simplicity we will consider the initial subsequence to be (n), but the arguments below go through for any initial subsequence.

Part (i). We prove only the results for \bar{S}^n and ρ as the other parts are similar. By Campbell's formula and Lemma 4.1 we have for a fixed $t \in [0,T]$, for all measurable functions $W_n(t,u,x)$: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}\left[\int_{0}^{t} \int_{\mathbb{R}} W_{n}(t, u, x) \mathcal{M}^{n, A}(du, dx)\right] = \int_{0}^{t} \int_{\mathbb{R}} W_{n}(t, u, x) n\lambda(u) g(x) du dx,$$

$$\mathbb{E}\left[\int_{0}^{t} \int_{\mathbb{R}} W_{n}(t, u, x) \mathcal{M}^{n, D}(du, dx)\right] = \int_{0}^{t} \int_{\mathbb{R}} W_{n}(t, u, x) n d_{n}(u) g(x) du dx. \tag{4.42}$$

Denote $\mathcal{M}_{*}^{n,A}$ and $\mathcal{M}_{*}^{n,D}$ to be the compensated random measures:

$$\mathcal{M}_{*}^{n,A} = \mathcal{M}^{n,A} - \mathcal{M}_{c}^{n,A}, \quad \mathcal{M}_{*}^{n,D} = \mathcal{M}^{n,D} - \mathcal{M}_{c}^{n,D},$$
 (4.43)

where $\mathcal{M}_{c}^{n,A}$ and $\mathcal{M}_{c}^{n,D}$ are as defined in (4.34).

Arrivals affecting number in service: We first investigate the stochastic integrals with respect to the random measure $\mathcal{M}^{n,A}$. By the decomposition (4.43) and Lemma 4.1, the first term in (4.13) becomes:

$$\bar{S}_t^{n,A} := X_t^{s,n,A} + Y_t^{s,n,A}, \tag{4.44}$$

where

$$X_t^{s,n,A} := \int_0^t \int_{\mathbb{R}} W_n^{s,A}(t,u,x) \mathcal{M}_*^{n,A}(du,dx), \quad \text{and} \quad Y_t^{s,n,A} := \int_0^t \mathbb{1}_{\{\bar{S}_{u-}^n < 1\}} \bar{G}(t-u) \lambda(u) du.$$

We can follow the same argument as in the proof of Proposition 3.1 to conclude similar to how we obtained (3.16) that for any subsequence (l_k) , there exists a subsubsequence $(r_k) \subset (l_k)$, such that for any $\phi \in L^1[0,T]$ there exists $w^1(u) \in L^{\infty}[0,T]$ and almost surely

$$\lim_{k \to \infty} \int_0^t \phi(u) \mathbb{1}_{\{\bar{S}_{u^-}^{r_k} < 1\}} du = \int_0^t \phi(u) w^1(u) du. \tag{4.45}$$

Furthermore, using similar arguments to how we obtained (3.25) we get almost surely

$$\bar{S}_t^{r_k,A} = X_t^{s,r_k,A} + Y_t^{s,r_k,A} \to \int_0^t w^1(u)\bar{G}(t-u)\lambda(u)du := Y^{s,A}. \tag{4.46}$$

in the uniform topology.

Departures affecting number in service: In this part we look at the stochastic integrals with respect to the random measure $\mathcal{M}^{n,D}$ in (4.13). Similar to (4.44) we have

$$\bar{S}_t^{n,D} = X_t^{s,n,D} + Y_t^{s,n,D}, \tag{4.47}$$

where

$$X_t^{s,n,D} = \int_0^t \int_{\mathbb{R}} W_n^{s,D}(t,u,x) \mathcal{M}_*^{n,D}(du,dx) \quad \text{and} \quad Y_t^{s,n,D} = \int_0^t \mathbb{1}_{\{\bar{Q}_{u-}^n > 0\}} \bar{G}(t-u) d_n(u) du. \quad (4.48)$$

We first analyze the term $Y^{s,n,D}$. Since the cumulative departures are upper bounded by the cumulative arrivals, by (4.48) and the integrability of λ we have

$$Y_t^{s,n,D} \le \int_0^t d_n(u)du \le \int_0^t \lambda(u)du < \infty. \tag{4.49}$$

Note that

$$Y_t^{s,n,D} - Y_s^{s,n,D} = \int_s^t \mathbb{1}_{\{\bar{Q}_{u-}^n > 0\}} \bar{G}(t-u) d_n(u) du + \int_0^s \mathbb{1}_{\{\bar{Q}_{u-}^n > 0\}} \left(\bar{G}(t-u) - \bar{G}(s-u) \right) d_n(u) du.$$

Since \bar{G} is non-increasing and bounded above by 1, we have

$$\sup_{n} \left| Y_t^{s,n,D} - Y_s^{s,n,D} \right| \le \sup_{n} \int_s^t d_n(u) du \le c_g(t-s) \int_0^T \lambda(u) du,$$

where the last inequality follows from (4.32). This Lipschitz continuity implies that $Y_t^{s,n,D}$ is equicontinuous. Therefore we have

$$\lim_{\delta \downarrow 0} \sup_{n} w'_{Y^{s,n,D}}(\delta) = 0. \tag{4.50}$$

By (4.49), (4.50), Theorem 2.1 and Prokhorov's theorem we can conclude that there exists $Y^{s,D} \in \mathbb{D}$ and a subsequence (n_k) such that almost surely

$$Y^{s,n_k,D} \xrightarrow{\mathbb{D}} Y^{s,D}$$
, almost surely. (4.51)

Since indicators are uniformly bounded, by [6, Thm 2.34], almost surely there exists a subsequence $(l_k) \subset (n_k)$ and $w^2(u) \in L^{\infty}[0, t]$, possibly depending on (l_k) , such that for any $\phi \in L^1[0, T]$

$$\lim_{k \to \infty} \int_0^t \phi(u) \mathbb{1}_{\{\bar{Q}_{u-}^{l_k} > 0\}} du = \int_0^t \phi(u) w^2(u) du, \quad \text{for all } t \in [0, T].$$
 (4.52)

Note that w^2 could still be random at this stage. In addition from Proposition 4.1.(iii) we have that there exists a bounded function d such that for any $\phi \in L^1[0,T]$ almost surely

$$\lim_{k \to \infty} \int_0^t \phi(u) dl_k(u) du = \int_0^t \phi(u) d(u) du, \quad \text{for all } t \in [0, T].$$

$$\tag{4.53}$$

Recall $Y^{s,n,D}$ from (4.48). By triangle inequality

$$\left| Y_t^{s,l_k,D} - \int_0^t w^2(u)\bar{G}(t-u)d(u)du \right| \\
\leq \left| \int_0^t \mathbb{1}_{\{\bar{Q}_{u-}^{l_k} > 0\}} \bar{G}(t-u)\left(d_{l_k}(u) - d(u)\right)du \right| + \left| \int_0^t \left(\mathbb{1}_{\{\bar{Q}_{u-}^{l_k} > 0\}} - w^2(u)\right)\bar{G}(t-u)d(u)du \right|, \quad (4.54)$$

where the right hand side converges to 0 as $k \to \infty$ almost surely, thanks to (4.52) and (4.53). Thus (4.54) yields for all $t \in [0, T]$, almost surely

$$\lim_{k \to \infty} Y_t^{s, l_k, D} = \int_0^t w^2(u) \bar{G}(t - u) d(u) du. \tag{4.55}$$

This means we can identify $Y^{s,D}$ in (4.51) from (4.55), that is:

$$Y_t^{s,D} = \int_0^t w^2(u)\bar{G}(t-u)d(u)du.$$
 (4.56)

This limiting function $Y^{s,D}$ is continuous because \bar{G} and w^2 are bounded, and d is integrable. It follows that the convergence in (4.51) is also under the uniform topology:

$$\lim_{k \to \infty} \sup_{t \in [0,T]} \left| Y_t^{s,l_k,D} - Y_t^{s,D} \right| = 0, \quad \text{almost surely.}$$
 (4.57)

Let us now analyze the term $X^{s,n,D}$. Similar to the argument in (3.20)-(3.21) one can also conclude that $X_t^{s,n,D}$ are uniformly bounded for all $t \in [0,T]$ and

$$\sup_{n} w'_{\bar{S}^{n,D}}(\delta) = 0. \tag{4.58}$$

Using (4.47), (4.58) and (4.50) we obtain

$$\lim_{\delta \downarrow 0} \sup_{n} w'_{X^{s,n,D}}(\delta) = 0. \tag{4.59}$$

The uniform boundedness and (4.59) together imply that $\{X_t^{s,n,D}\}_{n\geq 1}$ is tight. Consider the process

$$Z_{t}^{n} = \int_{0}^{t} \int_{\mathbb{R}} \mathcal{M}_{*}^{n,D}(du, dx) = \int_{0}^{t} \int_{\mathbb{R}} \mathcal{M}^{n,D}(du, dx) - \int_{0}^{t} \int_{\mathbb{R}} \mathcal{M}_{c}^{n,D}(du, dx). \tag{4.60}$$

Since $\int_0^t \int_{\mathbb{R}} \mathcal{M}_c^{n,D}(du,dx) = \int_0^t \int_{\mathbb{R}} d_n(u)g(x)dudx$ is a continuous function with bounded variation, by [32, Thm 26, Chapter 2] it has 0 quadratic variation. It follows that the quadratic variation of Z^n coincides with the quadratic variation of the pure jump process $\mathcal{M}^{n,D}([0,\cdot]\times\mathbb{R})$, i.e.

$$[Z^n, Z^n]_t = \sum_{i=1}^{\infty} (\mathbb{1}_{\{D_i \le t\}} \mathbb{1}_{\mathbb{R}}(V_{j_i}))^2 = \int_0^t \int_{\mathbb{R}} \mathcal{M}^{n, D}(du, dx).$$
 (4.61)

Consequently by [32, Thm 29, Chapter 2], and (4.48), (4.60) we have

$$[X^{s,n,D}, X^{s,n,D}]_t = \int_0^t \int_{\mathbb{R}} (W_n^{s,D}(t, u, x))^2 d[Z^n, Z^n]_t = \int_0^t \int_{\mathbb{R}} (W_n^{s,D}(t, u, x))^2 \mathcal{M}^{n,D}(du, dx).$$

By [32, Cor 3, Chapter 2] and (4.42) we conclude that

$$\mathbb{E}\left(X_T^{s,n,D}\right)^2 = \mathbb{E}\left(\left[X^{s,n,D}, X^{s,n,D}\right]_T\right) = \mathbb{E}\left[\int_0^T \int_{\mathbb{R}} \left(W_n^{s,D}(t,u,x)\right)^2 \mathcal{M}^{n,D}(du,dx)\right]$$

$$\leq \frac{1}{n} \int_0^T \int_{\mathbb{R}} g(x) d_n(u) du dx \leq \frac{1}{n} \int_0^T \lambda(u) du \to 0,$$

as $n \to \infty$, where we utilize (4.28) in the last inequality. Similar to the argument leading to (3.23) we obtained that

$$X^{s,n,D} \xrightarrow{p} 0$$
, in the uniform topology. (4.62)

From (4.62) we know that there exists a subsequence $(r_k) \subset (l_k)$ such that

$$\sup_{t \in [0,T]} |X^{s,r_k,D}| \to 0, \quad \text{almost surely.}$$
 (4.63)

For this sequence (r_k) we thus obtain from (4.47), (4.57) and (4.63) that almost surely

$$\bar{S}_{t}^{r_{k},D} = X_{t}^{s,r_{k},D} + Y_{t}^{s,r_{k},D} \to Y^{s,D}. \tag{4.64}$$

in the uniform topology, where the function $Y^{s,D}$ is identified by (4.56).

Conclusion: Let us denote for $t \in [0,T]$

$$\rho_t = Y_t^{s,A} + Y_t^{s,D} = \int_0^t w^1(u)\bar{G}(t-u)\lambda(u)du + \int_0^t w^2(u)\bar{G}(t-u)d(u)du.$$

Combining (4.46) and (4.64) we conclude that almost surely

$$\bar{S}^{r_k} = \bar{S}^{r_k,A} + \bar{S}^{r_k,D} \to \rho,$$

in the uniform topology, which is the desired convergence result for \bar{S}^n in (4.36). Number in buffer and departures: Similar arguments yield convergence of \bar{Q}^n and D^n in (4.36), in addition to the corresponding representations of the limits in (4.39) and (4.40). For the first term on the right hand side of (4.39), since $\mathbb{1}_{\{\bar{S}_{u-}^{r_k}=1\}} = 1 - \mathbb{1}_{\{\bar{S}_{u-}^{r_k}<1\}}$, by a diagonalization argument one can get for any $\phi \in L^1[0,T]$

$$\lim_{k \to \infty} \int_0^t \mathbb{1}_{\{\bar{S}^{r_k} = 1\}} \mathbb{1}_{\{\bar{Q}^{r_k} < b_{r_k}/r_k\}} \phi(u) = \int_0^t (1 - w^1(u)) w^3(u) \phi(u) du, \quad \text{almost surely.}$$

For the left hand side of (4.40), recall from (4.60) that

$$\bar{D}_t^n = \frac{1}{n} \int_0^t \int_{\mathbb{R}} \mathcal{M}_c^{n,D}(du, dx) + \frac{1}{n} Z_t^n = \int_0^t d_n(u) du + \frac{1}{n} Z_t^n.$$
 (4.65)

By Proposition 4.1 we know that for the subsequence (l_k) and d in (4.53), we have for $t \in [0, T]$

$$\int_0^t dl_k(u)du \to D_t = \int_0^t d(u)du,$$

in the uniform topology. By [32, Cor 3, Chapter 2], (4.42) and (4.61) we conclude that

$$\mathbb{E}\left(\frac{1}{n}Z_t^n\right)^2 = \frac{1}{n^2}\mathbb{E}([Z^n, Z^n]_t) = \frac{1}{n^2}\mathbb{E}\left[\int_0^t \int_{\mathbb{R}} \mathcal{M}^{n,D}(du, dx)\right]$$
$$\leq \frac{1}{n}\int_0^t \int_{\mathbb{R}} g(x)d_n(u)du\,dx \leq \frac{1}{n}\int_0^t \lambda(u)du \to 0.$$

By a similar argument leading to (4.63), we obtain that there exists a subsequence $(r_k) \subset (l_k)$ such that

$$\sup_{t \in [0,T]} \left| \frac{1}{r_k} Z_t^{r_k} \right| \to 0, \quad \text{almost surely.} \tag{4.66}$$

Combining (4.65)-(4.66) we can conclude that \bar{D}^{r_k} converge to $D_t = \int_0^t d(u)du$ almost surely in the uniform topology. This completes the proof of Part(i).

Part (ii). The weak* convergence of $\mathbb{1}_{\{\bar{Q}_{u}^{l_k}>0\}}$ has already been shown above in (4.52), and the counterpart of $\mathbb{1}_{\{\bar{S}_{v}^{l_k}<1\}}$ and $\mathbb{1}_{\{\bar{Q}_{v}^{l_k}< b_n/n\}}$ are similar.

Part (iii). The functions ρ, η, D have been identified in the proof of Part (i). It remains to show the constraints for functions w^1, w^2, w^3 . We now observe that the set $\{\bar{S}^n_{u-} < 1\}$ is identical to the set $\{\bar{S}^n_{u-} \le 1 - \frac{1}{n}\}$. This is because \bar{S}^n only takes values in $\{\frac{i}{n} : i = 1, \dots, n\}$. Therefore, (4.44) can be rewritten as

$$\bar{S}^{n,A}_t = X^{s,n,A}_t + \int_0^t \mathbb{1}_{\{\bar{S}^n_u = 1 - \frac{1}{n}\}} \bar{G}(t-u) \lambda(u) du.$$

Next, we try to find out w^1 . Recall the convergence stated in (4.36) in the uniform topology. Consequently fix $\varepsilon > 0$ and choose N large enough such that for all k > N we have $r_k > \frac{3}{\varepsilon}$, $\|\bar{S}^{r_k} - \rho\|_T < \frac{\varepsilon}{3}$ almost surely. Then it is readily checked that

$$\mathbb{1}_{\{\rho_{u-} \le 1 - \varepsilon\}} \le \mathbb{1}_{\{\bar{S}_{u-}^{r_k} \le 1 - 1/r_k\}} \le \mathbb{1}_{\{\rho_{u-} < 1 + \varepsilon\}}.$$
(4.67)

Therefore for any such that $\phi \in L^1[0,T]$ we have

$$\int_0^t \phi(u) \mathbb{1}_{\{\rho_{u-} \le 1 - \varepsilon\}} du \le \int_0^t \phi(u) \mathbb{1}_{\{\bar{S}_{u-}^{r_k} \le 1 - 1/r_k\}} du \le \int_0^t \phi(u) \mathbb{1}_{\{\rho_{u-} < 1 + \varepsilon\}} du, \quad \text{for all } t \in [0, T].$$

Note that $\lim_{\varepsilon \downarrow 0} \mathbb{1}_{\{\rho_{u-} < 1 - \varepsilon\}} = \mathbb{1}_{\{\rho_{u-} < 1\}}$ and $\lim_{\varepsilon \downarrow 0} \mathbb{1}_{\{\rho_{u-} < 1 + \varepsilon\}} = \mathbb{1}_{\{\rho_{u-} \leq 1\}} = 1$. Consequently taking $k \to \infty$ and then $\varepsilon \downarrow 0$ we have by the dominated convergence theorem and (4.45) that:

$$\int_0^t \phi(u) \mathbb{1}_{\{\rho_{u-} < 1\}} du \le \int_0^t w^1(u) \phi(u) du \le \int_0^t \phi(u) du, \quad \text{for all } t \in [0, T].$$

Since ϕ is arbitrary we have almost surely

$$\mathbb{1}_{\{\rho_{n-}<1\}} \le w^1(u) \le 1$$
, a.e. in $[0,T]$. (4.68)

Observe that one can also replace $\{\bar{Q}_{u-}^{r_k} < b_{r_k}/r_k\}$ by $\{\bar{Q}_{u-}^{r_k} \leq b_{r_k}/r_k - 1/r_k\}$ and $\{\bar{Q}_{u-}^{r_k} > 0\}$ by $\{\bar{Q}_{u-}^{r_k} \geq 1/r_k\}$. Similar to (4.67), for any $\varepsilon > 0$ one can choose large enough N such that for all k > N we have $r_k > \frac{3}{\varepsilon}$, $\|\bar{Q}^{r_k} - y\|_T < \frac{\varepsilon}{3}$ and $\|b_{r_k}/r_k - \beta\| < \frac{\varepsilon}{3}$ almost surely. Then it is readily checked that almost surely

$$\mathbb{1}_{\{\eta_{u-} \le \beta - \varepsilon\}} \le \mathbb{1}_{\{\bar{Q}_{u}^{r_k} < b_{r_k} / r_k - 1 / r_k\}} \le \mathbb{1}_{\{\eta_{u-} < \beta + \varepsilon\}},\tag{4.69}$$

$$\mathbb{1}_{\{\eta_{u-}>\varepsilon\}} \le \mathbb{1}_{\{\bar{Q}_{u}^{r_k} > 1/r_k\}} \le 1. \tag{4.70}$$

Similar to the argument leading to (4.68), from (4.70) we can conclude that almost surely for any $\phi \in L^1[0,T]$

$$\int_0^t \phi(u) \mathbb{1}_{\{\eta_{u-} > 0\}} du \le \int_0^t w^2(u) \phi(u) du \le \int_0^t \phi(u) du, \quad \text{for all } t \in [0, T],$$

and almost surely

$$\mathbb{1}_{\{\eta_{u-}>0\}} \le w^2(u) \le 1$$
, a.e. in $[0,T]$.

From (4.69) we can conclude that for any $\phi \in L^1[0,T]$

$$\mathbb{1}_{\{\eta_{u-}<\beta\}} \le \int_0^t w^3(u)\phi(u)du \le \int_0^t \phi(u)du,$$

and almost surely

$$\mathbb{1}_{\{\eta_{u-} < \beta\}} \le w^3(u) \le 1$$
, a.e. in $[0, T]$.

Therefore, by (4.38)-(4.40) and Definition 2.2 we conclude that $(\rho, \eta, d, w^1, w^2, w^3)$ is the solution to the discontinuous Volterra equation (4.41).

We have established a fluid limit for $(\bar{S}^n, \bar{Q}^n, \bar{D}^n)$ along a subsequence when the system starts empty. Now, we extend our considerations to a more general case.

Assumption 4.2. Let the conditions under Assumption 4.1 hold. In addition, let the number of customers in the system at time 0: N_0^n , satisfy the following asymptotic result:

$$\lim_{n \to \infty} \left| \frac{N_0^n}{n} - r_0 \right| = 0, \quad almost \ surely,$$

where $r_0 \in (0, 1 + \beta]$. Moreover, assume that the empirical distribution F^n of the remaining service times of the initial occupied servers satisfy

$$\lim_{n \to \infty} \sup_{t} |F^{n}(t) - F(t)| = 0, \quad almost \ surely$$

for some distribution F.

Proposition 4.3. Let Assumption 4.2 hold. Then

(i) For any T > 0 and for any subsequence of (n), there exists a further subsequence r_k and real-valued continuous, possibly stochastic processes ρ, η, D such that almost surely

$$\bar{S}_t^{r_k} \to \rho_t, \quad \bar{Q}_t^{r_k} \to \eta_t, \quad \bar{D}_t^{r_k} \to D_t,$$
 (4.71)

in the uniform topology.

(ii) Moreover, given (r_k) , almost surely there exist bounded, possibly stochastic processes w^1, w^2, w^3 such that

$$\mathbb{1}_{\{\bar{S}_{t-}^{r_k} < 1\}} \stackrel{*}{\rightharpoonup} w^1(t), \quad \mathbb{1}_{\{\bar{Q}_{t-}^{r_k} > 0\}} \stackrel{*}{\rightharpoonup} w^2(t), \quad \mathbb{1}_{\{\bar{Q}_{t-}^{r_k} < b_{r_k}/r_k\}} \stackrel{*}{\rightharpoonup} w^3(t), \quad in \ L^{\infty}[0, T].$$

$$(4.72)$$

(iii) Furthermore, almost surely, $(\rho, \eta, D, w^1, w^2, w^3)$ defined in (4.71)-(4.72) satisfy

$$\rho_t = \min\{r_0, 1\} \bar{F}(t) + \int_0^t w^1(u) \bar{G}(t - u) \lambda(u) du + \int_0^t w^2(u) \bar{G}(t - u) d(u) du, \tag{4.73}$$

$$\eta_t = \max\{r_0 - 1, 0\} + \int_0^t (1 - w^1(u))w^3(u)\lambda(u)du - \int_0^t w^2(u)d(u)du, \tag{4.74}$$

$$D_t = \min\{r_0, 1\}F(t) + \int_0^t w^1(u)G(t - u)\lambda(u)du + \int_0^t w^2(u)G(t - u)d(u)du, \tag{4.75}$$

and for almost every $t \in [0, T]$

$$\mathbb{1}_{\{\rho_t < 1\}} \le w^1(t) \le 1, \quad \mathbb{1}_{\{\eta_t > 0\}} \le w^2(t) \le 1, \quad \mathbb{1}_{\{\eta_t < \beta\}} \le w^3(t) \le 1.$$

That is, for almost all $\omega \in \Omega$ $(\rho(\omega), \eta(\omega), D, w^1(\omega), w^2(\omega), w^3(\omega))$ as in (4.73)-(4.75) is a solution, interpreted according to Definition 2.2, to the following non-linear discontinuous Volterra integral equation

$$\rho_{t} = \min\{r_{0}, 1\}\bar{F}(t) + \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}}\bar{G}(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}}\bar{G}(t-u)d(u)du,
\eta_{t} = \max\{r_{0} - 1, 0\} + \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}=1\}}\mathbb{1}_{\{\eta_{u-}<\beta\}}\lambda(u)du - \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}}d(u)du,
D_{t} = \min\{r_{0}, 1\}F(t) + \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}}G(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}}G(t-u)d(u)du.$$
(4.76)

Proof. At time 0, the number of customers in service is $\min\{N_0^n, n\}$ and the number of customers in buffer is $\max\{N_0^n - n, 0\}$. Let the remaining service times for the customers in service to be $(V_i^0)_{1 \le i \le \min\{N_0^n, n\}}$. Then, similar to (4.13)-(4.15) we have:

$$\begin{split} \bar{S}^n_t &= \frac{1}{n} \sum_{i=1}^{\min\{N^n_0,n\}} \mathbbm{1}_{\{V^0_i > t\}} + \int_0^t \int_{\mathbb{R}} W^{s,A}_n(t,u,x) \mathcal{M}^{n,A}(du,dx) + \int_0^t \int_{\mathbb{R}} W^{s,D}_n(t,u,x) \mathcal{M}^{n,D}(du,dx), \\ \bar{Q}^n_t &= \frac{1}{n} \max\{N^n_0 - n, 0\} + \int_0^t \int_{\mathbb{R}} W^{q,A}_n(t,u,x) \mathcal{M}^{n,A}(du,dx) - \int_0^t \int_{\mathbb{R}} W^{q,D}_n(t,u,x) \mathcal{M}^{n,D}(du,dx), \\ \bar{D}^n_t &= \frac{1}{n} \sum_{i=1}^{\min\{N^n_0,n\}} \mathbbm{1}_{\{V^0_i \le t\}} + \int_0^t \int_{\mathbb{R}} W^{d,A}_n(t,u,x) \mathcal{M}^{n,A}(du,dx) + \int_0^t \int_{\mathbb{R}} W^{d,D}_n(t,u,x) \mathcal{M}^{n,D}(du,dx). \end{split}$$

Observing that

$$\frac{1}{n} \sum_{i=1}^{\min\{N_0^n, n\}} \mathbb{1}_{\{V_i^0 > t\}} = \frac{\min\{N_0^n, n\}}{n} \sum_{i=1}^{\min\{N_0^n, n\}} \frac{\mathbb{1}_{\{V_i^0 > t\}}}{\min\{N_0^n, n\}}.$$

By Assumption 4.2 and (4.1) we have that

$$\lim_{n \to \infty} \sup_{t} \left| \frac{1}{n} \sum_{i=1}^{\min\{N_0^n, n\}} \mathbb{1}_{\{V_i^0 > t\}} - \min\{r_0, 1\} \bar{F}(t) \right| = 0, \quad \text{almost surely.}$$
 (4.77)

Similarly

$$\lim_{n \to \infty} \sup_{t} \left| \frac{1}{n} \sum_{i=1}^{\min\{N_0^n, n\}} \mathbb{1}_{\{V_i^0 \le t\}} - \min\{r_0, 1\} F(t) \right| = 0, \quad \text{almost surely}$$
 (4.78)

and obviously

$$\lim_{n \to \infty} \left| \frac{1}{n} \max\{N_0^n - n, 0\} - \max\{r_0 - 1, 0\} \right| = 0.$$
 (4.79)

Since we already analyzed the integration w.r.t $\mathcal{M}^{n,A}$ and $\mathcal{M}^{n,D}$ in Proposition 4.2, by (4.77)-(4.79) we get the desired results.

Now, we establish the existence of a unique (ρ, η, D) that satisfies (4.76) in the sense of Definition 2.2. Consequently, we obtain a unique fluid limit of the fraction of busy serverys, fraction of occupied buffers and the n-scaled cumulative departure rate.

Theorem 4.1. Let Assumption 4.2 hold. Then there exists a unique solution to the discontinuous Volterra integral equation (4.76), that is, there exist a unique solution (ρ, η, d) such that for $t \in [0, T]$

$$\rho_t = \min\{r_0, 1\} \bar{F}(t) + \int_0^t z^1(u) \bar{G}(t-u) \lambda(u) du + \int_0^t z^2(u) \bar{G}(t-u) d(u) du, \tag{4.80}$$

$$\eta_t = \max\{r_0 - 1, 0\} + \int_0^t (1 - z^1(u))z^3(u)\lambda(u)du - \int_0^t z^2(u)d(u)du, \tag{4.81}$$

$$D_t = \min\{r_0, 1\}F(t) + \int_0^t z^1(u)G(t-u)\lambda(u)du + \int_0^t z^2(u)G(t-u)d(u)du, \tag{4.82}$$

for some (z^1, z^2, z^3) that satisfies for almost every $t \in [0, T]$

$$\mathbb{1}_{\{\rho_t < 1\}} \le z^1(t) \le 1,\tag{4.83}$$

$$\mathbb{1}_{\{\eta_t > 0\}} \le z^2(t) \le 1,\tag{4.84}$$

$$\mathbb{1}_{\{n_t < \beta\}} \le z^3(t) \le 1, \quad and \tag{4.85}$$

$$0 \le \rho_t \le 1, \quad 0 \le \eta_t \le \beta, \quad \eta_t (1 - \rho_t) = 0.$$
 (4.86)

Proof. The existence of the solution directly follows from Proposition 4.3. Before we prove the uniqueness, let us first talk about the differentiability of the processes of interest. Since $\int_0^t d(u)du \le \int_0^t \lambda(u)du \le \infty$ for $t \in [0,T]$, $d \in L^1[0,T]$. For bounded x(t), since d(t), $\lambda(t)$, $g(t) \in L^1[0,T]$, by Young's convolution inequality we have

$$\frac{\partial}{\partial t}x(u)G(t-u)\lambda(u) = x(u)\lambda(u)g(t-u) \in L^1([0,T] \times [0,T]),$$

$$\frac{\partial}{\partial t}x(u)G(t-u)d(u) = x(u)d(u)g(t-u) \in L^1([0,T] \times [0,T]).$$

Therefore, by [35, Thm 2.7] we can differentiate both side of (4.82) for $t \in (0,T)$

$$d(t) = \min\{r_0, 1\} f(t) + \int_0^t \left(z^1(u)\lambda(u) + z^2(u)d(u)\right) g(t - u)du, \quad \text{a.e. in } [0, T].$$
 (4.87)

Differentiating both side of (4.80) and plugging in (4.87) we obtain

$$\rho_t' = -\min\{r_0, 1\} f(t) + \left(z^1(t)\lambda(t) + z^2(t)d(t)\right) - \int_0^t \left(z^1(u)\lambda(u) + z^2(u)d(u)\right) g(t - u)du$$

$$= \left(z^1(t)\lambda(t) + z^2(t)d(t)\right) - d(t), \quad \text{a.e. in } [0, T]. \tag{4.88}$$

By (4.86) we know that when $\eta_t > 0$, $\rho_t = 1$. Hence, in order to prove the uniqueness of (ρ, η, d) , we can divide the situations into four states:

- (1) $\rho_t < 1, \eta_t = 0.$
- (2) $\rho_t = 1, \eta_t = 0.$
- (3) $\rho_t = 1, 0 < \eta_t < \beta$.
- (4) $\rho_t = 1, \eta_t = \beta$.

Denote $\delta_i^k = \inf_{t > \gamma_i^{k-1}} \{t : (\rho_t, \eta_t) \in \text{state } i\}$ to be the k-th time (ρ_t, η_t) entering the i-th state, and $\gamma_i^k = \inf_{t > \delta_i^k} \{t : (\rho_t, \eta_t) \notin \text{state } i\}$ denote the k-th time (ρ_t, η_t) leaving the i-th state, i = 1, 2, 3, 4

and $k = 1, 2, 3, \dots$. With a similar argument after (3.35) we can conclude that there are at most countable many k. We discuss each state in the following for any $k \in \mathbb{N}$.

State 1: $\rho_t < 1, \eta_t = 0$. We need to identify ρ, d and γ_1^k in this state. From (4.83) we have for $t \in (\delta_1^k, \gamma_1^k), z^1(t) = 1$, a.e.. Plugging this into (4.81) we have

$$\eta_t = \eta_{\delta_1^k} - \int_{\delta_1^k}^t z^2(u) d(u) du.$$

Since η is continuous, for $t \in (\delta_1^k, \gamma_1^k)$ we have $\eta_t = \eta_{\delta_1^k} = 0$. Therefore, we can conclude that $z^2(t)d(t) = 0$, a.e.. Substituting z^1 and z^2d with their values in (4.80) and (4.87) we have for $t \in (\delta_1^k, \gamma_1^k)$

$$\rho_t = \min\{r_0, 1\} \bar{F}(t) + \int_0^{\delta_1^k} \left(z^1(u)\lambda(u) + z^2(u)d(u)\right) \bar{G}(t-u)du + \int_{\delta_1^k}^t \bar{G}(t-u)\lambda(u)du,$$

and

$$d(t) = \min\{r_0, 1\}f(t) + \int_0^{\delta_1^k} \left(z^1(u)\lambda(u) + z^2(u)d(u)\right)g(t-u)du + \int_{\delta_1^k}^t \lambda(u)g(t-u)du, \quad \text{a.e. in } [0, T].$$

We can see that $\gamma_1^k = \inf_{t > \delta_1^k} \{ \rho_t = 1 \}$ is unique if δ_1^k and $z^1(u)\lambda(u) + z^2(u)d(u)$ is known for $u \in [0, \delta_1^k)$. The next state can only be state 2.

State 2: $\rho_t = 1, \eta_t = 0$. We need to identify d and γ_2^k in this state. From (4.85) we have for $t \in (\delta_2^k, \gamma_2^k), z^3(t) = 1$, a.e.. Plugging this into (4.81) we have

$$\eta_t = \eta_{\delta_2^k} + \int_{\delta_2^k}^t (1 - z^1(u))\lambda(u)du - \int_{\delta_2^k}^t z^2(u)d(u)du.$$
 (4.89)

Since η is continuous, for $t \in (\delta_2^k, \gamma_2^k)$ we have $\eta_t = \eta_{\delta_2^k} = 0$. Therefore,

$$\int_{\delta_2^k}^t (1 - z^1(u))\lambda(u)du - \int_{\delta_2^k}^t z^2(u)d(u)du = 0.$$

Since t is arbitrary, we can conclude that for $t \in (\delta_2^k, \gamma_2^k)$

$$z^{1}(t)\lambda(t) + z^{2}(t)d(t) = \lambda(t)$$
, a.e. in [0, T]. (4.90)

Plugging (4.90) into (4.87) we get

$$d(t) = \min\{r_0, 1\} f(t) + \int_0^{\delta_2^k} \left(z^1(u)\lambda(u) + z^2(u)d(u)\right) g(t-u)du + \int_{\delta_2^k}^t \lambda(u)g(t-u)du, \quad \text{a.e. in } [0, T].$$

To identify γ_2^k we can plug (4.90) into (4.88). Since $\rho_t' = 0$ for $t \in (\delta_2^k, \gamma_2^k)$, we have

$$\rho'_t = \lambda(t) - d(t) = 0$$
, a.e. in $[0, T]$. (4.91)

Define

$$\delta_{2,1}^k = \sup_{t>\delta_2^k}\{t: \lambda(s) \geq d(s), \ \text{ for } a.e. \, s \in (\delta_2^k, t)\}$$

the first time after δ_2^k that $\lambda(t) < d(t)$ for a positive measure set, and

$$\delta_{2,3}^k = \sup_{t>\delta_2^k} \{t: \lambda(s) \leq d(s), \ \text{ for } a.e. \, s \in (\delta_2^k, t)\}$$

denote the first time after δ_2^k that $\lambda(t) > d(t)$ for a positive measure set. Since (4.91) is true for $t \in (\delta_2^k, \gamma_2^k)$, $\gamma_2^k = \min[\delta_{2,1}^k, \delta_{2,3}^k]$. If $\gamma_2^k = \delta_{2,1}^k$, the next state will be *state 1*. Indeed, since η_t

is continuous, there exist small enough $\varepsilon > 0$ s.t. for $t \in (\delta_{2,1}^k, \delta_{2,1}^k + \varepsilon)$, $0 \le \eta_t < \beta$ and thus $z^3(t) = 1$, a.e.. Consequently (4.89) is true in this interval. Applying Leibniz rule to (4.89) we have

$$\eta'_t = (1 - z^1(t))\lambda(t) - z^2(t)d(t), \quad \text{a.e. in } [0, T].$$
 (4.92)

Since $\eta_{\delta_{2,1}^k} = 0$ and $\eta_t \ge 0$ is continuous, there exist $\varepsilon' > 0$ such that $y'_t \ge 0$ for $t \in (\delta_{2,1}^k, \delta_{2,1}^k + \varepsilon')$. By (4.92) we have

$$z^{1}(t)\lambda(t) + z^{2}(t)d(t) \leq \lambda(t)$$
, a.e. in $[0, T]$.

By the definition of $\delta_{2,1}^k$ there exist a positive measure set $K \subset (\delta_{2,1}^k, \delta_{2,1}^k + \varepsilon')$ s.t. $\lambda(t) < d(t)$ for $t \in K$. Plugging this into the above inequality we have for $t \in K$

$$z^{1}(t)\lambda(t) + z^{2}(t)d(t) < d(t), \text{ a.e. in } [0,T].$$
 (4.93)

Therefore, by (4.88) and (4.93) we have

$$\rho'_t < 0$$
, for $a.e. t \in K$

By Fundamental Theorem of Calculus we have $\rho_t < 1$ for $t \in (\delta_{2,1}^k, \delta_{2,1}^k + \varepsilon')$, which is exactly state 1. Similarly, if $\gamma_2^k = \delta_{2,3}^k$, the next state will be *state 3*. We can see that γ_2^k is unique if δ_2^k and $z^1(u)\lambda(u) + z^2(u)d(u)$ is known for $u \in [0, \delta_2^k)$.

State 3: $\rho_t = 1, 0 < \eta_t < \beta$. We need to identify y, d and γ_3^k in this state. From (4.84) and (4.85) we know that for $t \in (\delta_3^k, \gamma_3^k)$

$$z^{2}(t) = 1, z^{3}(t) = 1,$$
 a.e. in $[0, T]$. (4.94)

Plugging (4.94) into (4.88), we have

$$\rho'_t = z^1(t)\lambda(t) \ge 0$$
, a.e. in $[0, T]$.

Since $\rho_t \leq 1$, we have $\rho_t' = 0$ for $t \in (\delta_3^k, \gamma_3^k)$. Consequently

$$z^{1}(t)\lambda(t) = 0$$
, for a.e. $t \in (\delta_{3}^{k}, \gamma_{3}^{k})$ (4.95)

Therefore, plugging (4.94)-(4.95) into (4.87) we obtain for almost every $t \in [0, T]$

$$d(t) = \min\{r_0, 1\} f(t) + \int_0^{\delta_3^k} \left(z^1(u)\lambda(u) + z^2(u)d(u)\right) g(t - u)du + \int_{\delta_3^k}^t d(u)g(t - u)du, \quad (4.96)$$

By [7, Thm 6.3.1] there exist a unique solution d(t) to the Volterra integral equation (4.96) for $t \in (\delta_3^k, \gamma_3^k)$. With this solution and (4.94), (4.95) we can obtain

$$\eta_t = \eta_{\delta_3^k} + \int_{\delta_5^k}^t \lambda(u) - d(u)du.$$

Define $\delta_{3,2}^k = \inf_{t>\delta_3^k} \{t: \eta_t = 0\}$ and $\delta_{3,4}^k = \inf_{t>\delta_3^k} \{t: \eta_t = \beta\}$. Then $\gamma_3^k = \min[\delta_{3,2}^k, \delta_{3,4}^k]$. If $\gamma_3^k = \delta_{3,2}^k$, the next state will be state 2. If $\gamma_3^k = \delta_{3,4}^k$, the next state will be state 4. We can see that γ_3^k is unique if δ_3^k , $\eta_{\delta_3^k}$ and $z^1(u)\lambda(u) + z^2(u)d(u)$ is known for $u \in [0, \delta_3^k)$.

State 4: $\rho_t = 1, \eta_t = \beta$. We need to identify d and γ_4^k in this state. From (4.84) we have for $t \in (\delta_4^k, \gamma_4^k), z^2(t) = 1$, a.e.. Similar to the argument leading to (4.95) we obtain

$$z^{1}(t)\lambda(t) = 0$$
 for a.e. $t \in (\delta_{4}^{k}, \gamma_{4}^{k}).$ (4.97)

Substituting $z^1\lambda$ and z^2 with their values in (4.87) we have for almost every $t \in [0,T]$

$$d(t) = \min\{r_0, 1\} f(t) + \int_0^{\delta_4^k} \left(z^1(u)\lambda(u) + z^2(u)d(u) \right) g(t-u)du + \int_{\delta_4^k}^t d(u)g(t-u)du,$$
 (4.98)

By [7, Thm 6.3.1] again there exist a unique solution d(t) to the Volterra integral equation (4.98) for $t \in (\delta_4^k, \gamma_4^k)$. Plugging this solution, (4.97) and $z^2(t) = 1$ into (4.81) we have

$$\beta = \eta_{\delta_4^k} + \int_{\delta_4^k}^t z^3(u)\lambda(u) - d(u)du.$$
 (4.99)

Differentiating both side of (4.99) we get

$$z^3(t)\lambda(t) = d(t)$$
 for a.e. $t \in (\delta_4^k, \gamma_4^k)$.

It is easy to see that when $d(t) > \lambda(t)$, there does not exist $z^3(t)$ satisfies (4.85). Therefore,

$$\gamma_4^k = \sup_{t > \delta_4^k} \{t : d(s) \le \lambda(s), \text{ for a.e. } s \in (\delta_4^k, t)\}.$$

We can see that γ_4^k is unique if δ_4^k and $z^1(u)\lambda(u) + z^2(u)d(u)$ is known for $u \in [0, \delta_4^k)$. The next state can only be *State 3*.

Note that in every state above one can obtain unique $(\rho_t, \eta_t, d(t))$. Additionally, in every state above one can obtain almost surely either $(z^1(t), z^2(t))$ or $z^1(u)\lambda(u) + z^2(u)d(u)$ and thus unique γ_i^k , i = 1, 2, 3, 4. Since k is arbitrary, we construct a unique solution $(\rho_t, \eta_t, d(t))$ to the system (4.80)-(4.86). Indeed, if there exist two different solution $\rho_t^1, \eta_{t,1}, d_1(t)$ and $\rho_t^2, \eta_{t,2}, d_2(t)$ satisfying (4.80)-(4.82), the first time they differ must be one of those δ_i^k or γ_i^k . However, this violates the uniqueness established above in each state and leads to a contradiction. Therefore, we obtain the uniqueness of ρ, η, d and the resulting solution satisfies (4.80)-(4.82) and

$$\begin{cases} State \ 1, & z^1(t) = 1, z^2(t)d(t) = 0, z^3(t) = 1, \\ State \ 2, & z^1(t)\lambda(t) + z^2(t)d(t) = \lambda(t), z^3(t) = 1, \\ State \ 3, & z^1(t)\lambda(t) = 0, z^2(t) = 1, z^3(t) = 1, \\ State \ 4, & z^1(t)\lambda(t) = 0, z^2(t) = 1, z^3(t)\lambda(t) = d(t). \end{cases}$$

Similar to Theorem 3.2, we now provide possible solutions to the auxiliary functions (z^1, z^2, z^3) .

Theorem 4.2. Under the setting of Theorem 4.2, the solution to (4.80)-(4.86) satisfies

$$\begin{cases}
z^{1}(t) = 1, z^{2}(t)d(t) = 0, z^{3}(t) = 1, & \rho_{t} < 1, \eta_{t} = 0 \\
z^{1}(t)\lambda(t) + z^{2}(t)d(t) = \lambda(t), z^{3}(t) = 1, & \rho_{t} = 1, \eta_{t} = 0 \\
z^{1}(t)\lambda(t) = 0, z^{2}(t) = 1, z^{3}(t) = 1, & \rho_{t} = 1, 0 < \eta_{t} < \beta \\
z^{1}(t)\lambda(t) = 0, z^{2}(t) = 1, z^{3}(t)\lambda(t) = d(t), & \rho_{t} = 1, \eta_{t} = \beta
\end{cases}$$

$$(4.100)$$

In particular, the tuple $(\rho, \eta, d, z^1, z^2, z^3)$ satisfying (4.80)-(4.85) and

$$\begin{cases} z^1(t) = 1, z^2(t) = 0, z^3(t) = 1, & \rho_t < 1, \eta_t = 0 \\ z^1(t) = 1, z^2(t) = 0, z^3(t) = 1, & \rho_t = 1, \eta_t = 0 \\ z^1(t) = 0, z^2(t) = 1, z^3(t) = 1, & \rho_t = 1, 0 < \eta_t < \beta \\ z^1(t) = 0, z^2(t) = 1, z^3(t) = d(t)/\lambda(t) \land 1, & \rho_t = 1, \eta_t = \beta \end{cases}$$

is a solution to the system (4.80)-(4.86). Moreover, the functions $z^1\lambda + z^2d$ and $z^3\lambda$ are unique almost surely.

Proof. From the proof of Theorem 4.1 we have (z^1, z^2, z^3) satisfy

$$\begin{cases} \textit{State 1}, & z^1(t) = 1, z^2(t)d(t) = 0, z^3(t) = 1, \\ \textit{State 2}, & z^1(t)\lambda(t) + z^2(t)d(t) = \lambda(t), z^3(t) = 1, \\ \textit{State 3}, & z^1(t)\lambda(t) = 0, z^2(t) = 1, z^3(t) = 1, \\ \textit{State 4}, & z^1(t)\lambda(t) = 0, z^2(t) = 1, z^3(t) = d(t)/\lambda(t) \wedge 1. \end{cases}$$

It is easy to see that $z^1\lambda + z^2d$ and $z^3\lambda$ are unique almost surely. Choosing $z^1 = 1, z^2 = 0$ in state 2 and, $z^1 = 0$ in state 3 and 4, we get our desired result.

Theorem 4.3. Let Assumption 4.2 hold. Then

(i) For any T > 0, there exist real-valued continuous deterministic processes ρ, η, D such that almost surely

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left| \bar{S}^n_t - \rho_t \right| = 0, \quad \lim_{n\to\infty} \sup_{t\in[0,T]} \left| \bar{Q}^n_t - \eta_t \right| = 0, \quad \lim_{n\to\infty} \sup_{t\in[0,T]} \left| \bar{D}^n_t - D_t \right| = 0. \tag{4.101}$$

(ii) Moreover, there exist bounded functions w^1, w^2, w^3 such that almost surely

$$\mathbb{1}_{\{\bar{S}_{t-}^{n}<1\}}\lambda(t) + \mathbb{1}_{\{\bar{Q}_{t-}^{n}>0\}}d(t) \stackrel{*}{\rightharpoonup} w^{1}(t)\lambda(t) + w^{2}(t)d(t), \quad and$$

$$\mathbb{1}_{\{\bar{Q}_{t-}^{n}$$

where w^1, w^2, w^3 satisfy (4.100).

(iii) Furthermore, $(\rho, \eta, D, w^1, w^2, w^3)$ defined in (4.101)-(4.102) satisfy

$$\rho_t = \min\{r_0, 1\} \bar{F}(t) + \int_0^t w^1(u) \bar{G}(t-u) \lambda(u) du + \int_0^t w^2(u) \bar{G}(t-u) d(u) du, \tag{4.103}$$

$$\eta_t = \max\{r_0 - 1, 0\} + \int_0^t (1 - w^1(u))w^3(u)\lambda(u)du - \int_0^t w^2(u)d(u)du, \tag{4.104}$$

$$D_t = \min\{r_0, 1\} F(t) + \int_0^t w^1(u) G(t - u) \lambda(u) du + \int_0^t w^2(u) G(t - u) d(u) du, \tag{4.105}$$

and for almost every $t \in [0, T]$

$$\mathbb{1}_{\{\rho_t < 1\}} \le w^1(t) \le 1, \quad \mathbb{1}_{\{\eta_t > 0\}} \le w^2(t) \le 1, \quad \mathbb{1}_{\{\eta_t < \beta\}} \le w^3(t) \le 1.$$

That is, $(\rho, \eta, d, w^1, w^2, w^3)$ as in (4.103)-(4.105) is a solution, interpreted according to Definition 2.2, to the following non-linear discontinuous Volterra integral equation

$$\rho_t = \min\{r_0, 1\} \bar{F}(t) + \int_0^t \mathbb{1}_{\{\rho_{u-} < 1\}} \bar{G}(t-u) \lambda(u) du + \int_0^t \mathbb{1}_{\{\eta_{u-} > 0\}} \bar{G}(t-u) d(u) du, \tag{4.106}$$

$$\eta_t = \max\{r_0 - 1, 0\} + \int_0^t \mathbb{1}_{\{\rho_{u-} = 1\}} \mathbb{1}_{\{\eta_{u-} < \beta\}} \lambda(u) du - \int_0^t \mathbb{1}_{\{\eta_{u-} > 0\}} d(u) du, \tag{4.107}$$

$$D_{t} = \min\{r_{0}, 1\}F(t) + \int_{0}^{t} \mathbb{1}_{\{\rho_{u-}<1\}}G(t-u)\lambda(u)du + \int_{0}^{t} \mathbb{1}_{\{\eta_{u-}>0\}}G(t-u)d(u)du.$$
 (4.108)

Proof. Part (i). From Proposition 4.3, for any subsequence there exists a further subsequence (r_k) such that almost surely

$$\bar{S}_t^{r_k} \to \rho_t, \quad \bar{Q}_t^{r_k} \to \eta_t, \quad \bar{D}_t^{r_k} \to D_t,$$

in the uniform topology, where (ρ, η, D) solve (4.106)-(4.108) path by path. By Theorem 4.1, (ρ, η, D) is unique. Consequently, the limiting functions (ρ, η, D) are deterministic. Moreover, by the uniqueness of (ρ, η, D) again we can conclude that the entire sequence $(\bar{S}^n, \bar{Q}^n, \bar{D}^n)$ converges to (ρ, η, D) almost surely in the uniform topology.

Part (ii). By Proposition 4.3 we have for every subsequence there exists a subsubsequence (r_k) and bounded, possibly stochastic processes w^1, w^2, w^3 such that almost surely

$$\mathbb{1}_{\{\bar{S}^{r_k}_{t-} < 1\}} \stackrel{*}{\rightharpoonup} w^1(t), \quad \mathbb{1}_{\{\bar{Q}^{r_k}_{t-} > 0\}} \stackrel{*}{\rightharpoonup} w^2(t), \quad \mathbb{1}_{\{\bar{Q}^{r_k}_{t-} < b_{r_k}/r_k\}} \stackrel{*}{\rightharpoonup} w^3(t), \quad \text{in } L^{\infty}[0, T].$$

Consequently,

$$\mathbb{1}_{\{\bar{S}_{t-}^{r_k} < 1\}} \lambda(t) + \mathbb{1}_{\{\bar{Q}_{t-}^{r_k} > 0\}} d(t) \stackrel{*}{\rightharpoonup} w^1(t) \lambda(t) + w^2(t) d(t), \text{ and}$$

$$\mathbb{1}_{\{\bar{Q}_{t-}^{r_k} < b_{r_k} / r_k\}} \lambda(t) \stackrel{*}{\rightharpoonup} w^3(t) \lambda(t), \text{ in } L^{\infty}[0, T].$$
(4.109)

By Theorem 4.2 we know that $w^1\lambda + w^2d$ and $w^3\lambda$ is unique. Therefore the weak-star convergence in (4.109) hold for the entire sequence.

Part (iii). This follows directly from Proposition 4.3.(iii) and Theorem 4.1-4.2.

Similar to Corollary 3.1, an asymptotic result of the acceptance probability can be obtained. We state the following result without proof.

Corollary 4.1. The acceptance probability of the n-th $M_t/G/n/n+b_n$ model $\mathbb{P}(\bar{Q}_{t-}^n<\frac{b_n}{n})$ satisfies the following convergence

$$\mathbb{P}\left(\bar{Q}_{u-}^n < \frac{b_n}{n}\right) \to w^3(u), \quad \text{for } \lambda\text{-almost every } u \in [0, T],$$

where w^3 is defined in Theorem 4.3.

Remark 4.1. Similar to Remark 3.5, the function w^3 can be discontinuous even when λ is continuous.

5. Numerics and Operational Perspectives

5.1. Numerical Methods for Discontinuous VIE. This section outlines a simple procedure to numerically solve the discontinuous VIEs (3.50) and (4.106)-(4.108), using an explicit Euler discretization. The main computational challenge lies in updating the auxiliary functions z or (z^1, z^2, z^3) in tandem with the solution trajectories ρ or (ρ, η, d) at each iteration. Algorithm 1 details the steps to solve (3.50) following the solution framework described in Theorems 3.1-3.2, while Algorithm 2 extends this to the coupled system (4.106)-(4.108) using Theorems 4.1 - 4.2.

Algorithm 1 VIE for Zero-Buffer Loss System

```
1: Input: Initial value \rho_{t_0}, time points t_0, t_1, \ldots, t_N, functions f, g, \bar{F}, \bar{G}, \lambda.
2: Initialization: Set z(t_0) = 0.
3: for i = 0 to N - 1 do
                                                                       \triangleright Determine z values for time t_{i+1} based on state at t_i
4:
          if \rho_{t_i} < 1 then
5:
               z_{t_{i+1}} \leftarrow 1
          else if \rho_{t_i} = 1 then
               z_{t_{i+1}} \leftarrow \min \left( \frac{1}{\lambda(t_i)} \left( \rho_0 f(t_i) + \sum_{j=1}^i z(t_j) \lambda(t_j) g(t_i - t_j) \right), 1 \right)
7:
8:

ho Update state for time t_{i+1} by discretizing integral equations \rho_{t_{i+1}} \leftarrow \rho_0 \bar{F}(t_{i+1}) + \sum_{j=1}^{i+1} z(t_j) \lambda(t_j) \bar{G}(t_{i+1} - t_j)
```

9:
$$\rho_{t_{i+1}} \leftarrow \rho_0 F(t_{i+1}) + \sum_{j=1}^{i+1} z(t_j) \lambda(t_j) G(t_{i+1} - t_j)$$

10:

11: Output: The sequence of values for ρ and z.

Algorithm 2 VIE for Loss System with Buffer

```
1: Input: Initial value r_0 or (\rho_{t_0}, \eta_{t_0}), threshold \beta, time points t_0, t_1, \ldots, t_N, functions f, g, \bar{F}, \bar{G}, \lambda.
2: Initialization: Set (z_{t_0}^1, z_{t_0}^2, z_{t_0}^3, d(t_0)) = (0, 0, 0, 0).
3: for i = 0 to N - 1 do
                                                                                                           \triangleright Determine z values for time t_{i+1} based on state at t_i
              if \rho_{t_i} < 1 and \eta_{t_i} = 0 then (z_{t_{i+1}}^1, z_{t_{i+1}}^2, z_{t_{i+1}}^3) \leftarrow (1, 0, 1) else if \rho_{t_i} = 1 and \eta_{t_i} = 0 then
4:
5:
6:
              else if \rho_{t_i} = 1 and \eta_{t_i} = 0 then (z_{t_{i+1}}^1, z_{t_{i+1}}^2, z_{t_{i+1}}^3) \leftarrow (1, 0, 1) else if \rho_{t_i} = 1 and 0 < \eta_{t_i} < \beta then (z_{t_{i+1}}^1, z_{t_{i+1}}^2, z_{t_{i+1}}^3) \leftarrow (0, 1, 1) else if \rho_{t_i} = 1 and \eta_{t_i} = \beta then
7:
8:
9:
```

```
11: (z_{t_{i+1}}^1, z_{t_{i+1}}^2) \leftarrow (0, 1)

12: z_{t_{i+1}}^3 \leftarrow \min(d(t_i)/\lambda(t_i), 1)

13: end if

Dupdate state for time t_{i+1} by discretizing integral equations

14: d_{t_{i+1}} \leftarrow \min(r_0, 1) f(t_{i+1}) + \sum_{j=1}^{i+1} \left(z^1(t_j)\lambda(t_j) + z^2(t_j)d(t_j)\right) g(t_{i+1} - t_j)

15: \rho_{t_{i+1}} \leftarrow \min(r_0, 1) \bar{F}(t_{i+1}) + \sum_{j=1}^{i+1} \left(z^1(t_j)\lambda(t_j) + z^2(t_j)d(t_j)\right) \bar{G}(t_{i+1} - t_j)

16: \eta_{t_{i+1}} \leftarrow \max(r_0 - 1, 0) + \sum_{j=1}^{i+1} \left[(1 - z^1(t_j))z^3(t_j)\lambda(t_j) - z^2(t_j)d(t_j)\right]

17: end for

18: Output: The sequences of values for (\rho, \eta, d) and (z^1, z^2, z^3).
```

5.1.1. Example 1: Zero-buffer Loss System. We first solve the VIE (3.29) using Algorithm 1. The simulated system has n=150 servers and Lognormal(-0.5,2) service times. Two types of arrival rates are used: periodic with $\lambda(t)=2/3(1+\sin(2\pi t/10))$ as in Figure 3a, and episodic with $\lambda(t)=0.005 \cdot t(T-t)$ as in Figure 4a. The simulated trajectory \bar{N}^n closely matches the fluid-limit solution ρ , as in Figures 3b and 4b, confirming the convergence in Theorem 3.3. Repeating the simulation over R=200 replications shows that the empirical blocking probability $\bar{B}^n(t)$ aligns well with the theoretical 1-w(t), supporting Corollary 3.1. Figures 3c-3d and 4c-4d illustrate this relationship for n=150 and n=5000 for both arrival types.

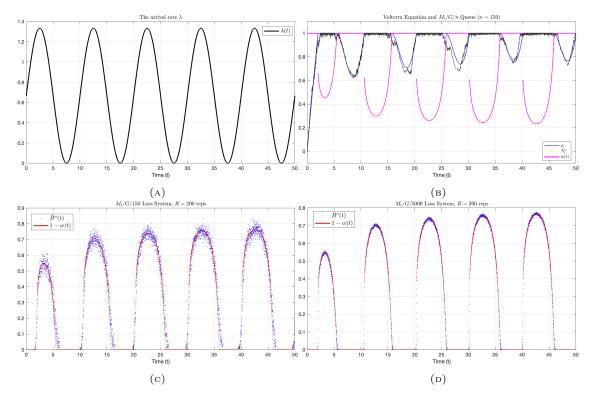


Figure 3. Zero-Buffer Loss Queue with Periodic Arrival Rate

5.1.2. Example 2: Finite-buffer Loss System. Next, we solve the VIE system (4.76) using Algorithm 2. Again, the system has n=150 servers and Lognormal(-0.5, 1.2) service times, with two types of arrival rates: periodic with $\lambda(t)=\frac{2}{3}(1.5+\sin(\frac{2\pi t}{10}))$ as in Figure 5a and periodic with $\lambda(t)=0.005 \cdot t(T-t)$ as in Figure 6a. The simulated trajectories \bar{S}^n and \bar{Q}^n track the theoretical

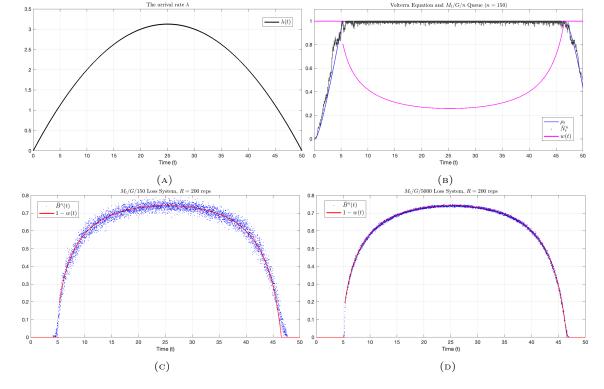


Figure 4. Zero-Buffer Loss Queue with Episodic Arrival Rate

 (ρ, η) closely, as in Figures 5b and 6b, confirming the convergence in Theorem 4.3. Similarly, the blocking probability $\bar{B}^n(t)$ aligns with $1 - w^3(t)$, validating the finite-buffer fluid approximation in Corollary 4.1. This is illustrated in Figures 5c-5d and 6c-6d for n=150 and n=5000 both arrival types.

Remark 5.1. As noted in Remark 3.5 and 4.1, the auxiliary functions w and w^3 may exhibit discontinuities. In the numerical results, this discontinuity becomes evident as system size n increases (e.g., n = 5000). For smaller systems, the blocking probability appears smoother, but the underlying discontinuity emerges clearly in the large-system limit.

5.2. Operational Perspectives. Finite capacity is a defining characteristic of many real-world service systems, such as call centers, emergency departments, and cloud resource pools. Such systems exhibit non-stationary queuing dynamics due to time-varying arrivals and general service times, making accurate transient analysis crucial for operational insights.

Our fluid limits for the zero-buffer $M_t/G/n/n$ and finite-buffer $M_t/G/n/n + b_n$ provide first-order approximations of the system occupancy, queue length, departure process, and acceptance probabilities as $n \to \infty$. These deterministic approximations offer a tractable foundation for operational optimization: they allow one to compute time-varying blocking probabilities directly and to optimize server and buffer capacities against transient performance constraints.

5.2.1. Staffing Optimization in Zero-Buffer Systems. Consider a sequence of non-stationary $M_t/G/c_n/c_n$ loss systems, with $c_n = \lfloor nc \rfloor$ servers and arrivals satisfying Assumption 3.2. For simplicity, assume that the system starts empty. Let the scaled number in the system or the proportion of occupied servers in the n-th system be \bar{N}_t^n and the n-scaled cumulative departure process be \bar{D}_t^n . Then similar to the treatise done in Section 3 and Theorem 3.2, we let the patient

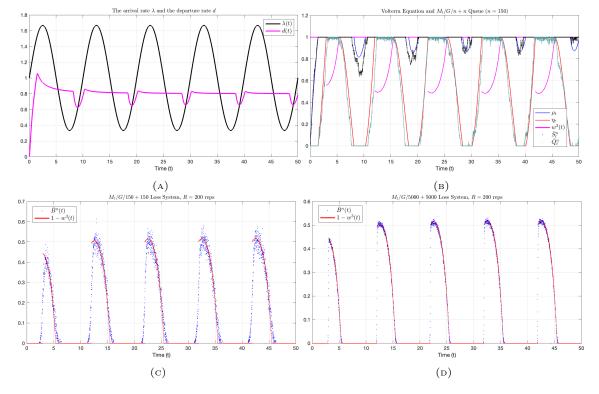


FIGURE 5. Finite Buffer Loss Queue with Periodic Arrival Rate

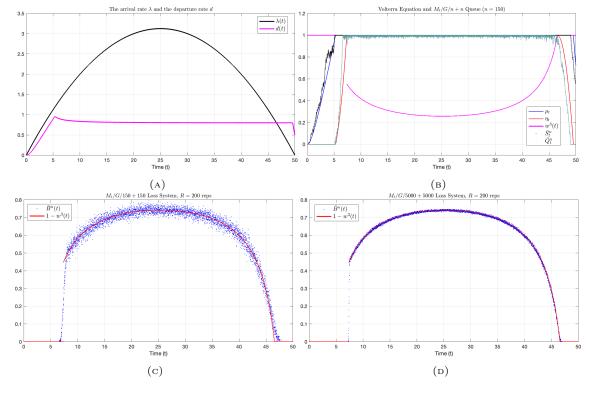


FIGURE 6. Finite Buffer Loss Queue with Episodic Arrival Rate

reader work out the details to conclude that

$$\lim_{t \to \infty} \sup_{t \in [0,T]} \left| \bar{N}_t^n - \rho_t \right| = 0, \quad \lim_{t \to \infty} \sup_{t \in [0,T]} \left| \bar{D}_t^n - D_t \right| = 0,$$

almost surely where D(t) is the fluid cumulative departure rate given by $D_t = \int_0^t d(u)du$ and $d(\cdot)$ is the fluid instantaneous departure rate whose dynamics is presented below. In addition, ρ_t solves the discontinuous non-linear VIE given by

$$\rho_t = \int_0^t \mathbb{1}_{\{\rho_u - \langle c\}} \bar{G}(t - u) \lambda_u du. \tag{5.1}$$

We note that the number of servers could be time-varying with $c_n(t) = \lfloor nc(t) \rfloor$ in which case our results will remain valid as long as the capacity constraint $\mathbb{1}_{\{\rho_u < c_u\}}$ is incorporated in (5.1). However for simplicity we consider $c(\cdot)$ to be constant. Under Definition 2.2, (5.1) can be equivalently expressed as

$$\rho_t = \int_0^t w_c(u)\bar{G}(t-u)\lambda(u)du, \tag{5.2}$$

where

$$w_c(t) = \begin{cases} 1, & \text{if } \rho_t < c, \\ \frac{d(t)}{\lambda(t)} \wedge 1, & \text{if } \rho_t = c, \end{cases}$$
 (5.3)

and

$$d(t) = \int_0^t w_c(u)g(t-u)\lambda(u)du. \tag{5.4}$$

Furthermore the acceptance probability converges uniformly

$$\lim_{n \to \infty} \sup_{t \in [0,T]} |P(\bar{N}_t^n < c_n) - w_c(t)| = 0.$$
 (5.5)

Equations (5.2)-(5.4) and the convergence result (5.5) direct server capacity optimization while maintaining the transient blocking probability above a given threshold. This is particularly relevant in emergency departments or call centers where it is important from a managerial perspective to minimize the total number of blocked patients or customers. Summarizing, we solve the following problem in the fluid limit

$$\min c$$
, such that $\inf_{t \in [0,T]} w_c(t) \ge 1 - \alpha$,

where α is the maximum allowable instantaneous blocking probability. Since the infimum of the acceptance probability w increases as the capacity c increases, the problem admits a unique optimal solution to the above constrained optimization problem. Therefore we can apply standard root-finding techniques (e.g. the bisection method) to obtain the optimal server capacity c^* . For each value of c, we solve the discontinuous VIE by Algorithm 1 to get $\inf_{t \in [0,T]} w_c(t)$, and stop searching once $\inf_{t \in [0,T]} w_{c^*}(t) = 1 - \alpha$. Numerical results are presented in Figures 7 and 8, respectively, for periodic and episodic arrival rates, and for two choices of α .

5.2.2. Joint Staffing and Buffer Capacity Optimization. Now consider a sequence of non-stationary $M_t/G/c_n/c_n + b_n$ loss queuing systems, where $c_n = \lfloor nc \rfloor$, buffer size $b_n = \lfloor n\beta \rfloor$, and arrivals satisfy Assumption 4.1. Similar to the previous case, server and buffer size could be time-varying with $c_n(t) = \lfloor nc(t) \rfloor$ and $b_n(t) = \lfloor n\beta(t) \rfloor$, and our results would still hold valid as long as $c(\cdot)$ and $\beta(\cdot)$ are piecewise constant. However, for simplicity, we do not consider those generalizations and also assume that the system starts empty. Let the scaled number being served or the proportion of occupied servers be \bar{N}_t^n , the scaled number waiting in buffer be \bar{Q}_t^n and the n-scaled cumulative

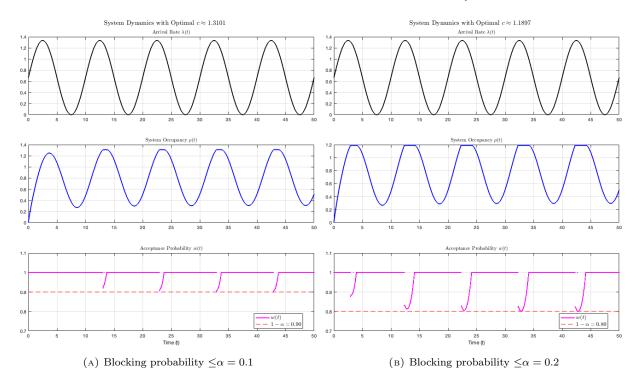
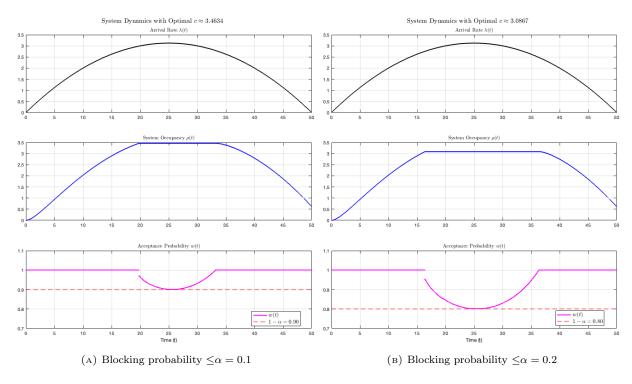


Figure 7. Optimal server capacity in zero-buffer loss queue with periodic arrival rate.



 $\label{eq:Figure 8. Optimal server capacity in zero-buffer loss queue with episodic arrival rate.$

departure process be \bar{D}_t^n . Then similar to the treatise done in Section 4 and Theorem 4.3, we let

the patient reader work out the details to conclude that

$$\lim_{t\to\infty} \sup_{t\in[0,T]} \left| \bar{N}_t^n - \rho_t \right| = 0, \quad \lim_{t\to\infty} \sup_{t\in[0,T]} \left| \bar{Q}_t^n - \eta_t \right| = 0 \quad \lim_{t\to\infty} \sup_{t\in[0,T]} \left| \bar{D}_t^n - D_t \right| = 0,$$

almost surely where D(t) is the fluid cumulative departure rate given by $D_t = \int_0^t d(u)du$ and $d(\cdot)$ is the fluid instantaneous departure rate whose dynamics is presented below. In addition, the fluid limits (ρ, η, d) solves a coupled discontinuous nonlinear VIE system which interpreted according to Definition 2.2 reads

$$\rho_{t} = \int_{0}^{t} w_{c,\beta}^{1}(u) \bar{G}(t-u) \lambda(u) du + \int_{0}^{t} w_{c,\beta}^{2}(u) \bar{G}(t-u) d(u) du,
\eta_{t} = \int_{0}^{t} (1 - w_{c,\beta}^{1}(u)) w_{c,\beta}^{3}(u) \lambda(u) du - \int_{0}^{t} w_{c,\beta}^{2}(u) d(u) du,
D_{t} = \int_{0}^{t} w_{c,\beta}^{1}(u) G(t-u) \lambda(u) du + \int_{0}^{t} w_{c,\beta}^{2}(u) G(t-u) d(u) du,$$
(5.6)

where the auxiliary functions $w_{c,\beta}^1$, $w_{c,\beta}^2$, $w_{c,\beta}^3$ evolve similar to (5.7):

$$w_{c,\beta}^{1}(t) = 1, w_{c,\beta}^{2}(t) = 0, w_{c,\beta}^{3}(t) = 1, \qquad \rho_{t} < c, \eta_{t} = 0$$

$$w_{c,\beta}^{1}(t) = 1, w_{c,\beta}^{2}(t) = 0, w_{c,\beta}^{3}(t) = 1, \qquad \rho_{t} = c, \eta_{t} = 0$$

$$w_{c,\beta}^{1}(t) = 0, w_{c,\beta}^{2}(t) = 1, w_{c,\beta}^{3}(t) = 1, \qquad \rho_{t} = c, 0 < \eta_{t} < \beta$$

$$w_{c,\beta}^{1}(t) = 0, w_{c,\beta}^{2}(t) = 1, w_{c,\beta}^{3}(t) = d(t)/\lambda(t) \wedge 1, \qquad \rho_{t} = c, \eta_{t} = \beta$$

$$(5.7)$$

The acceptance probability again satisfies

$$\lim_{n \to \infty} \sup_{t \in [0,T]} |P(\bar{Q}_t^n < b_n) - w_{c,\beta}^3(t)| = 0.$$
 (5.8)

Using equations (5.6)-(5.3) and the convergence result (5.8), we can formulate a joint staffing-buffer optimization problem in the fluid limit constrained to maintain the transient blocking probability above a threshold:

$$\min v \cdot c + (1-v) \cdot \beta$$
, such that $\inf_{t \in [0,T]} w_{c,\beta}^3(t) \ge 1-\alpha$,

where v weights the relative cost of servers and buffer space, and α denotes the maximum allowable instantaneous blocking probability. For each c, the infimum of the acceptance probability w^3 increases with buffer size β . Thus, there exists a unique β_c such that $\inf_{t\in[0,T]}w_{c,\beta_c}^3(t)=1-\alpha$. We can perform a grid search for c and apply standard root-finding techniques (e.g., bisection method) to determine the optimal β_c for each c. At each c and β , we solve the discontinuous VIE system using Algorithm 2 to obtain $\inf_{t\in[0,T]}w_{c,\beta}^3(t)$. The search terminates when $\inf_{t\in[0,T]}w_{c,\beta_c}^3(t)=1-\alpha$. The optimal solution is the capacity c^* that minimizes $v\cdot c^*+(1-v)\cdot \beta_{c^*}$ during the grid search. Numerical results are presented in Figure 9. It is important to note that this solution provides a rudimentary approach to solving the constrained optimization problem. While there may be more effective optimization techniques available, the focus of this paper is not on exploring such alternative solutions.

6. Conclusion

This paper developed a unified fluid-limit framework for nonstationary many-server loss systems with general service-time distributions. In the first part, we established a functional strong law of large numbers for the zero-buffer $M_t/G/n/n$ model via a discontinuous Volterra integral equation representation. The second part extended this analysis to the finite-buffer $M_t/G/n/(n+b_n)$ model, showing that the joint dynamics of servers, buffer occupancy, and departures satisfy a coupled

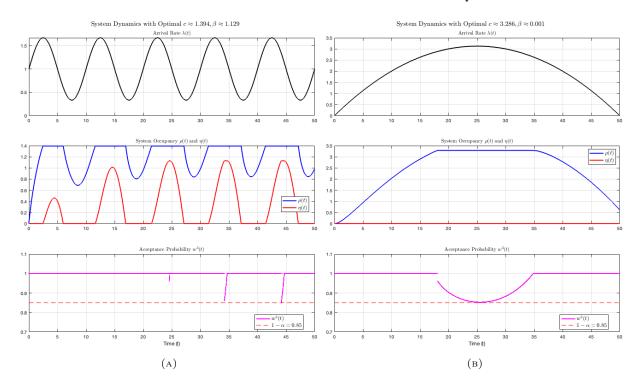


Figure 9. Optimal server and buffer capacity with periodic and episodic arrival rates.

system of discontinuous Volterra equations. In both regimes, we proved existence and uniqueness of the limiting trajectories and convergence of the associated time-varying acceptance and blocking probabilities.

The results demonstrate that deterministic fluid models can accurately describe transient behavior in large-scale, non-Markovian, time-varying loss systems. The discontinuous Volterra structure captures admission control and boundary effects within a mathematically rigorous and computationally tractable framework, bridging the gap between asymptotic theory and operational approximation.

Beyond theoretical insight, the model provides a practical basis for performance evaluation and real-time decision-making. We showed how the fluid limit can be used for optimal staffing and buffer capacity design, and the same structure can naturally extend to dynamic control settings. Future work may pursue diffusion refinements, stochastic perturbation analysis, and optimization-based control formulations, further integrating transient queuing dynamics into the broader landscape of stochastic operations management.

References

- [1] Anders Reenberg Andersen, Bo Friis Nielsen, and Line Blander Reinhardt. Optimization of hospital ward resources with patient relocation using markov chain modeling. *European Journal of Operational Research*, 260(3):1152–1163, 2017.
- [2] René Bekker, Ger Koole, and Dennis Roubos. Flexible bed allocations for hospital wards. *Health Care Management Science*, 20(4):453–466, 2016.
- [3] Patrick Billingsley. Convergence of probability measures. John Wiley & Sons, 2013.
- [4] Patrick Billingsley. Probability and measure. John Wiley & Sons, 2017.
- [5] Pierre Brémaud. An introduction to applied probability. Springer, 2024.
- [6] Alberto Bressan. Lecture notes on functional analysis. American Mathematical Society, 2012.
- [7] Hermann Brunner. Volterra integral equations: an introduction to theory and applications, volume 30. Cambridge University Press, 2017.
- [8] Prakash Chakraborty and Harsha Honnappa. A many-server functional strong law for a non-stationary loss model. Operations Research Letters, 49(3):338–344, 2021.

- [9] A. M. de Bruin, R. Bekker, L. van Zanten, and G. M. Koole. Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2009.
- [10] Rick Durrett. Probability: theory and examples, volume 49. Cambridge university press, 2019.
- [11] B. Eklundh. Channel utilization and blocking probability in a cellular mobile telephone system with directed retry. *IEEE Transactions on Communications*, 34(4):329–337, 1986.
- [12] Agner Krarup Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197, 1917.
- [13] Gerald B Folland. Real analysis: modern techniques and their applications. John Wiley & Sons, 1999.
- [14] Linda Green and Peter Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. Management Science, 37(1):84–97, 1991.
- [15] Linda V Green, Peter J Kolesar, and Ward Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.
- [16] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations* research, 29(3):567–588, 1981.
- [17] Robert C. Hampshire, Shan Bao, Walter S. Lasecki, Andrew Daw, and Jamol Pender. Beyond safety drivers: Applying air traffic control principles to support the deployment of driverless vehicles. PLOS ONE, 15(5):e0232837, May 2020.
- [18] Daehyoung Hong and S.S. Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology*, 35(3):77–92, 1986.
- [19] Otis B Jennings, Avishai Mandelbaum, William A Massey, and Ward Whitt. Server staffing to meet time-varying demand. Management Science, 42(10):1383–1394, 1996.
- [20] Haya Kaspi and Kavita Ramanan. Law of large numbers limits for many-server queues. *The Annals of Applied Probability*, 21(1):33–114, 2011.
- [21] F. P. Kelly. Blocking probabilities in large circuit-switched networks. Advances in Applied Probability, 18(2):473–505, 1986.
- [22] T. Kiffe. A discontinuous volterra integral equation. Journal of Integral Equations, 1(3):193–200, 1979.
- [23] T. Kiffe and M. Stecher. Existence and uniqueness of solutions to abstract volterra integral equations. Proceedings of the American Mathematical Society, 68(2):169–175, 1978.
- [24] T. Kiffe and M. Stecher. L² solutions of volterra integral equations. SIAM Journal on Mathematical Analysis, 10(2):274–280, 1979.
- [25] V.O.K. Li, Wanjiun Liao, Xiaoxin Qiu, and E.W.M. Wong. Performance model of interactive video-on-demand systems. IEEE Journal on Selected Areas in Communications, 14(6):1099–1109, 1996.
- [26] Yunan Liu and Ward Whitt. A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. Operations Research Letters, 40(5):307–312, 2012.
- [27] Yunan Liu and Ward Whitt. Many-server heavy-traffic limit for queues with time-varying parameters. *The Annals of Applied Probability*, 24(1):378–421, 2014.
- [28] Avi Mandelbaum, William A Massey, and Martin I Reiman. Strong approximations for markovian service networks. Queueing Systems, 30(1):149–201, 1998.
- [29] William A. Massey and Ward Whitt. An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *The Annals of Applied Probability*, 4(4):1145–1160, 1994.
- [30] Jamol Pender. Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering* and Informational Sciences, 29(1):27–49, 2015.
- [31] Jamol Pender and Young Myoung Ko. Approximations for the queue length distributions of time-varying many-server queues. *INFORMS Journal on Computing*, 29(4):688–704, 2017.
- [32] Phillip E Protter. Stochastic integration and differential equations, 2004.
- [33] Josh Reed. The G/GI/N queue in the Halfin–Whitt regime. The Annals of Applied Probability, 19(6):2211–2269, 2009.
- [34] Sidney I Resnick. Extreme values, regular variation, and point processes, volume 4. Springer Science & Business Media, 2008.
- [35] Jiří Šremr. On differentiation of a lebesgue integral with respect to a parameter. *Mathematics for Applications*, 1(1):91–116, 2012.
- [36] Shahin Vakilinia, Mustafa Mehmet Ali, and Dongyu Qiu. Modeling of the resource allocation in cloud computing centers. *Computer Networks*, 91:453–470, 2015.
- [37] Meiqian Wang, Shuo Li, Eric W. M. Wong, and Moshe Zukerman. Performance analysis of circuit switched multi-service multi-rate networks with alternative routing. *Journal of Lightwave Technology*, 32(2):179–200, 2014.
- [38] Ward Whitt. Time-varying queues. Queueing models and service management, 1(2), 2018.
- [39] Ward Whitt and Wei You. Time-varying robust queueing. Operations Research, 67(6):1766–1782, 2019.

- [40] Ward Whitt and Jingtong Zhao. Many-server loss models with non-poisson time-varying arrivals. *Naval Research Logistics (NRL)*, 64(3):177–202, 2017.
- [41] A. Zalesky, H.L. Vu, Z. Rosberg, E.W.M. Wong, and M. Zukerman. Obs contention resolution performance. *Performance Evaluation*, 64(4):357–373, 2007.
- [42] Jiheng Zhang. Fluid models of many-server queues with abandonment. Queueing Systems, 73(2):147–193, 2013.
- (M. Wang) Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802 United States, *Email address*: mvw5822@psu.edu
- (P. Chakraborty) Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802 United States, *Email address*: prakashc@psu.edu